UNIVERSITAT ROVIRA I VIRGILI
DESARROLLO DE DIFERENTES MÉTODOS DE SELECCIÓN
DE VARIABLES PARA SISTEMAS MULTISENTORIALES
Oscar Eduardo Gualdron Guerrero
ISBN:978-84-693-4070-7/DL:T-1167-2010

Desarrollo de diferentes métodos de selección de variables para sistemas multisensoriales

Autor:
Oscar Eduardo Gualdrón Guerrero

Directores:
Dr. Eduard Llobet Valero
Dr. Jesús Brezmes Llecha

Escola Tècnica Superior D' Enginyeria
Departament D`Enginyeria Electrònica Elèctrica I Automàtica
Universitat Rovira I Virgili
Tarragona (España), 12 de Septiembre de 2006

UNIVERSITAT ROVIRA I VIRGILI
DESARROLLO DE DIFERENTES MÉTODOS DE SELECCIÓN
DE VARIABLES PARA SISTEMAS MULTISENTORIALES
Oscar Eduardo Gualdron Guerrero
ISBN:978-84-693-4070-7/DL:T-1167-2010

UNIVERSITAT ROVIRA I VIRGILI
DESARROLLO DE DIFERENTES MÉTODOS DE SELECCIÓN
DE VARIABLES PARA SISTEMAS MULTISENTORIALES
Oscar Eduardo Gualdron Guerrero
ISBN:978-84-693-4070-7/DL:T-1167-2010

Dedicado a:

Mis Padres

Luis A. Gualdrón. y Cruz Delia Guerrero.

Mis hermanos y Sobrino

Andrés, Daniel, Julián, Luis Alberto y
Andrés David

UNIVERSITAT ROVIRA I VIRGILI
DESARROLLO DE DIFERENTES MÉTODOS DE SELECCIÓN
DE VARIABLES PARA SISTEMAS MULTISENTORIALES
Oscar Eduardo Gualdron Guerrero
ISBN:978-84-693-4070-7/DL:T-1167-2010

ISBN:978-84-693-4070-7/DL:T-1167-2010

Índice

ISBN:978-84-693-4070-7/DL:T-1167-2010

UNIVERSITAT ROVIRA I VIRGILI
DESARROLLO DE DIFERENTES MÉTODOS DE SELECCIÓN
DE VARIABLES PARA SISTEMAS MULTISENTORIALES
Oscar Eduardo Gualdron Guerrero
ISBN:978-84-693-4070-7/DL:T-1167-2010

Prefacio

UNIVERSITAT ROVIRA I VIRGILI
DESARROLLO DE DIFERENTES MÉTODOS DE SELECCIÓN
DE VARIABLES PARA SISTEMAS MULTISENTORIALES
Oscar Eduardo Gualdron Guerrero
ISBN:978-84-693-4070-7/DL:T-1167-2010

ISBN:978-84-693-4070-7/DL:T-1167-2010

"Con la paciencia y la tranquilidad se logra todo...y algo más"

Benjamín Franklin

ISBN:978-84-693-4070-7/DL:T-1167-2010

Agradecimientos

Quiero empezar expresando mis más profundos agradecimientos a la Universidad Rovira I Virgili y a la Universidad de Pamplona (Colombia) por permitir realizar mis estudios Doctorales en España y especialmente en Tarragona. Durante estos cuatro largos años, no sólo he tenido la oportunidad de adquirir conocimiento científico y formarme como un futuro investigador, sino que también he podido hacerlo en mi aspecto personal y profesional.

En el tiempo que he realizado mis estudios he conocido gente muy valiosa que me han brindado su conocimiento. En especial quiero agradecer a mis dos directores de tesis, los Doctores Eduard Llobet Valero y Jesús Brezmes Llecha, un privilegio que muy pocas personas pueden tener ya que me han brindado no sólo su conocimiento, sino también su confianza y apoyo incondicional para poder alcanzar este logro.

Por otro lado, más que un agradecimiento quiero dedicar esta tesis a mi numerosa y especial familia, principalmente a las personas con las cuales Dios me ha bendecido y permitido formar parte de ellos, como son mis padres Luis Alberto y Cruz Delia, se que sin ellos no hubiera logrado ser la persona que soy, “gracias por dedicarme parte de su vida, por guiarme en el buen camino y darme todo su apoyo incondicional”, A mis hermanos Andrés, Daniel, Julián, Luis Alberto y a mi sobrinito especial Andrés David por estar siempre pendientes de mí en todo momento y por su constante ánimo y apoyo.

A Mary, que a pesar de la distancia siempre ha estado a mi lado y que fue muchas veces mi fuente de inspiración y de alegría. “Gracias por tu paciencia y por el amor que me brindas”.

Finalmente, agradecer a mi larga y valiosa lista de amigos, no los nombraré porque ya saben que seguramente me pasaré a alguno, sólo decir que siempre los llevaré en mi corazón por todo lo que me han brindado en este tiempo, por el apoyo y ánimo en los momentos difíciles y por las alegrías que me hicieron vivir, “Gracias por todo amigos”.

ISBN:978-84-693-4070-7/DL:T-1167-2010

Resumen

Los sistemas de olfato electrónico son instrumentos que han sido desarrollados para emular a los sistemas de olfato biológicos. A este tipo de ingenios se les ha conocido popularmente como narices electrónicas (NE). Los científicos e ingenieros que siguen perfeccionando este tipo de instrumento trabajan en diferentes frentes, como son el del desarrollo de nuevos sensores de gases (con mejor discriminación y mayor sensibilidad), el de la adaptación de técnicas analíticas como la espectrometría de masas (MS) en substitución de la tradicional matriz de sensores químicos, la extracción de nuevos parámetros de la respuesta de los sensores (preprocesado) o incluso en el desarrollo de técnicas más sofisticadas para el procesado de datos.

Uno de los principales inconvenientes que en la actualidad presentan los sistemas de olfato artificial es la alta dimensionalidad de los conjuntos a analizar, debido a la gran cantidad de parámetros que se obtienen de cada medida. El principal objetivo de esta tesis ha sido estudiar y desarrollar nuevos métodos de selección de variables con el fin de reducir la dimensionalidad de los datos y así poder optimizar los procesos de reconocimiento en sistemas de olfato electrónico basados en sensores de gases o en espectrometría de masas.

Para poder evaluar la importancia de los métodos y comprobar si ayudan realmente a solucionar la problemática de la dimensionalidad se han utilizado cuatro conjuntos de datos pertenecientes a aplicaciones reales que nos permitieron comprobar y comparar los diferentes métodos implementados de forma objetiva. Estos cuatro conjuntos de datos se han utilizado en tres estudios cuyas conclusiones repasamos a continuación:

En el primero de los estudios se ha demostrado que diferentes métodos (secuenciales o estocásticos) pueden ser acoplados a clasificadores fuzzy ARTMAP o PNN y ser usados para la selección de variables en problemas de análisis de gases en sistemas multisensoriales. Los métodos fueron aplicados simultáneamente para identificar y cuantificar tres compuestos orgánicos volátiles y sus mezclas binarias construyendo sus respectivos modelos neuronales de clasificación.

ISBN:978-84-693-4070-7/DL:T-1167-2010

El segundo trabajo que se incluye en esta tesis propone una nueva estrategia para la selección de variables que se ha mostrado eficaz ante diferentes conjuntos de datos provenientes de sistemas olfativos basados en espectrometría de masas (MS). La estrategia ha sido aplicada inicialmente a un conjunto de datos consistente de mezclas sintéticas de compuestos volátiles. Este conjunto ha sido usado para mostrar que el proceso de selección es viable para identificar un mínimo número de fragmentos que permiten la discriminación correcta entre mezclas usando clasificadores fuzzy ARTMAP. Además, dada la naturaleza simple del problema planteado, fue posible mostrar que los fragmentos seleccionados, son fragmentos de ionización característicos de las especies presentes en las mezclas a ser discriminadas. Una vez demostrado el correcto funcionamiento de esta estrategia, se aplicó esta metodología a otros dos conjuntos de datos (aceite de oliva y jamones ibéricos, respectivamente).

El tercer estudio tratado en esta tesis ha girado en torno al desarrollo de un nuevo método de selección de variables inspirado en la concatenación de varios procesos de "backward selection". El método está especialmente diseñado para trabajar con Support Vector machines (SVM) en problemas de clasificación o de regresión. La utilidad del método ha sido evaluada usando dos de los conjuntos de datos ya utilizados anteriormente.

Como conclusión se puede decir que para los diferentes conjuntos estudiados, la inclusión de un proceso previo de selección de variables da como resultado una reducción drástica en la dimensionalidad y un aumento significativo en los correspondientes resultados de clasificación. Los métodos introducidos aquí no solo son útiles para resolver problemas de narices electrónicas basadas en MS, sino también para cualquier aplicación de sistemas de olfato artificial que presenten problemas de alta dimensionalidad como en el caso de los conjuntos de datos estudiados en este trabajo.

UNIVERSITAT ROVIRA I VIRGILI
DESARROLLO DE DIFERENTES MÉTODOS DE SELECCIÓN
DE VARIABLES PARA SISTEMAS MULTISENTORIALES
Oscar Eduardo Gualdron Guerrero
ISBN:978-84-693-4070-7/DL:T-1167-2010

1.

Introducción

UNIVERSITAT ROVIRA I VIRGILI
DESARROLLO DE DIFERENTES MÉTODOS DE SELECCIÓN
DE VARIABLES PARA SISTEMAS MULTISENTORIALES
Oscar Eduardo Gualdron Guerrero
ISBN:978-84-693-4070-7/DL:T-1167-2010

1.1 Motivación

Los Sistemas de Olfato Electrónico (SDOE) son instrumentos que han sido desarrollados intentando emular el funcionamiento de los sistemas de olfato biológicos. A este tipo de ingenios se les ha conocido popularmente como Narices Electrónicas (NE). Los científicos e ingenieros que siguen perfeccionando este tipo de sistemas trabajan en diferentes frentes para intentar neutralizar las limitaciones prácticas que presentan. Entre las estrategias bajo estudio, se pueden destacar las siguientes:

- Desarrollo de nuevas tecnologías de fabricación y de síntesis de materiales que permitan desarrollar sensores de gases más sensibles y con un mayor poder de discriminación.
- Incorporación de nuevas técnicas de detección como la espectrometría de masas en substitución de la tradicional matriz de sensores químicos.
- Diseño de nuevos métodos de extracción de información mediante la incorporación de nuevos parámetros y métodos de operación.
- Desarrollo de técnicas sofisticadas de procesado de datos y reconocimiento de patrones.

De todo ello se deriva que en la mayoría de intentos por mejorar este tipo de instrumentos se puede encontrar la tendencia común a generar un mayor número de descriptores (parámetros, variables) por cada medida realizada, lo cual, lejos de ser beneficioso, genera nuevos problemas a resolver.

Por ejemplo, en una NE con N sensores, el mínimo número de parámetros extraídos en cada medida será N (uno por sensor) aunque pueden ser muchos más cuando utilizamos información dinámica. La obtención de un amplio número de variables descriptoras por experimento (resultado de multiplicar el número de sensores por el número de parámetros) puede a priori parecer deseable, pero probablemente no todos los descriptores sean relevantes para las tareas de clasificación y cuantificación encomendadas.

Por tanto, con las nuevas tendencias se hace cada vez más evidente la necesidad de aplicar un método de selección de variables que permita eliminar del conjunto de descriptores aquellos que sean redundantes o que sólo introduzcan ruido al sistema de reconocimiento. A juicio de un buen número de investigadores [1-11] este proceso es una de las claves para mejorar la precisión de los sistemas de olfato electrónico en su difícil tarea de analizar aromas simples o complejos.

A modo de ejemplo, podemos comentar las razones que aconsejan realizar una selección de las variables a utilizar por el algoritmo de reconocimiento de patrones en los sistemas de olfato electrónico basados en matrices de sensores:

- Los parámetros que provienen de sensores poco sensibles a los compuestos volátiles de interés pueden presentar una alta varianza que no relacionada a cambios en la composición del aroma analizado sino debida al ruido, lo que no solo no ayuda en la tarea de reconocimiento de estos volátiles sino que dificulta dicho proceso.
- Los parámetros que provengan de sensores que tengan un comportamiento muy parecido en la detección de determinados compuestos volátiles proporcionarán información redundante. La información redundante aumenta la complejidad del sistema sin aportar mejoras substanciales en el modelo de aprendizaje, lo cual empeora la capacidad de generalización del equipo.
- Como regla general, la utilización de un número elevado de sensores en la matriz de detección incrementará el tamaño, el peso y costo del sistema final (por lo tanto, influenciará negativamente en sus características comerciales). Idealmente, el número de sensores debe minimizarse siempre y cuando no se comprometa el funcionamiento del sistema de NE.

Reducir el número de variables con el fin de optimizar el funcionamiento implica un cierto riesgo de pérdida de información. Por ese motivo las variables deben seleccionarse cuidadosamente. Una selección inadecuada de variables puede llevar a un funcionamiento inaceptable del sistema. La figura 1.1 resume las posibles situaciones en las que nos podemos encontrar, señalando que opción sería la más deseable para el sistema.

ISBN:978-84-693-4070-7/DL:T-1167-2010

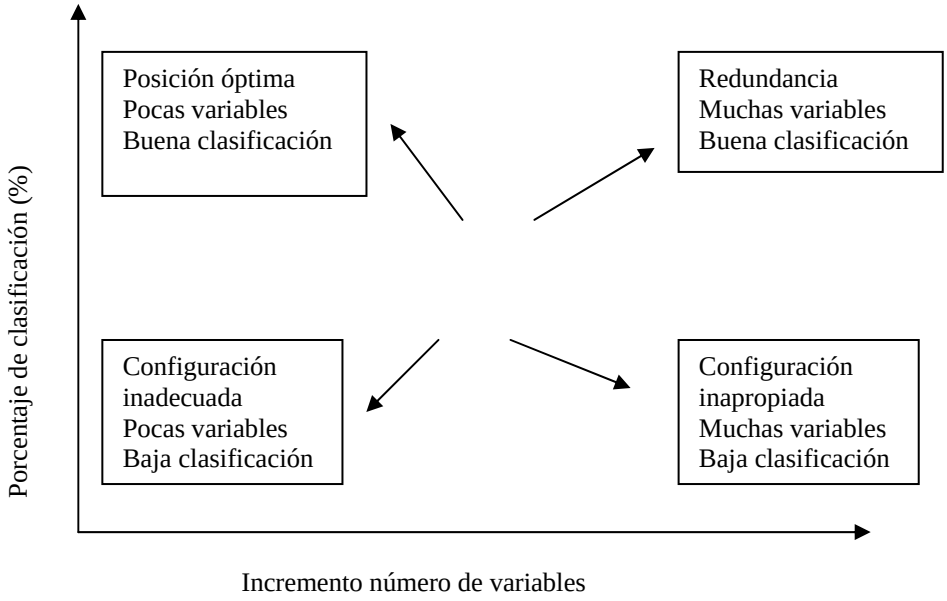

Figura 1.1 configuración de la selección de variables.

De todas formas, la determinación de la combinación óptima de variables no es una tarea trivial. Una exhaustiva búsqueda de todas las combinaciones posibles es computacionalmente costosa debido al elevado número de configuraciones que se pueden formar a partir de un conjunto de *n* elementos.

Considerando que se tiene una matriz de *n* variables diferentes y que se desea determinar el número de combinaciones diferentes *N* de *p* posibles sensores, ignorando *p*=0, el número de posibles combinaciones viene dado por la ecuación:

$$N = \sum_{p=2}^{n} \frac{n!}{(n-p)!\,p!} \qquad n \geq p \qquad (1.1)$$

El problema de la selección de variables se vuelve todavía mucho más importante en el caso de los sistemas de olfato electrónico basados en espectrometría de masas, configuración que ha presentado resultados muy prometedores en problemas de

clasificación y predicción de olores y aromas [12-19]. Las NE basadas en espectrometría de masas consideran a cada relación masa-carga (m/z) como un sensor diferente, por lo que se trata de un instrumento con una matriz de tantos sensores como variables masa-carga m/z se incluyan en el rango de análisis. Es importante destacar que en muchos casos, los mejores resultados se obtienen con un número muy pequeño de relaciones m/z (provenientes de iones fragmentados característicos de una aplicación).

El uso indiscriminado de la totalidad del espectro de variables m/z puede conllevar a introducir ruido no deseado y empobrecer los resultados de clasificación/predicción obtenidos. La elección óptima de las relaciones m/z que describan mejor la aplicación que se busca es de gran importancia para que el sistema de NE funcione correctamente.

Muchos autores han propuesto diferentes estrategias para seleccionar la configuración óptima del sistema de olfato electrónico basándose en el conocimiento previo de los analitos presentes en cada aplicación [20-25]. Este conocimiento previo normalmente se obtiene mediante técnicas tradicionales como la cromatografía de gases- espectrometría de masas (GC-MS). Es cierto que mediante un método que optimice la separación cromatográfica se pueden obtener los mismos o mejores resultados que mediante la espectrometría directa. Sin embargo, la complicada puesta a punto de dicho método (que requiere de personal cualificado), su complicada ejecución y tardanza hacen de la espectrometría directa un rival sin igual, siempre y cuando sus resultados sean los adecuados.

En definitiva, los métodos de selección de variables son necesarios para conseguir que la espectrometría directa de masas sea útil, y si no se dispone de personal especializado estos métodos deben realizar automáticamente la selección de las relaciones m/z que son relevantes para la aplicación diseñada sin necesidad de conocer previamente la naturaleza exacta de los analitos

ISBN:978-84-693-4070-7/DL:T-1167-2010

1.2 Objetivos

Una vez se ha identificado la problemática de la alta dimensionalidad en los datos y la necesidad de buscar conjuntos reducidos de variables que permitan optimizar el proceso de reconocimiento, se plantea como objetivo principal de esta tesis el desarrollo de nuevos métodos de selección de variables basados tanto en modelos secuenciales como estocásticos acoplados con modelos predictivos basados en diferentes redes neuronales (fuzzy ARTMAP, PNN) y métodos de reconocimiento de patrones como los Support Vector Machines (SVM).

Para la evaluación de todas estas posibilidades se utilizarán una amplia variedad de conjuntos de datos reales para la comprobación y comparación de los respectivos métodos implementados. Con todo este trabajo se persigue reducir la dimensionalidad de los problemas de identificación en aplicaciones de sistemas de olfato electrónico.

El objetivo principal de la tesis puede ser subdividido en diferentes etapas sobre los que se fundamenta esta memoria:

- Desarrollo de métodos para la selección de variables basados en técnicas secuenciales como son *forward selection, backward elimination y stepwise selection*)
- Desarrollo de métodos para la selección de variables basados en técnicas estocásticas como los algoritmos genéticos (GA) y el *simulated annealing* (SA).
- Determinación de los diferentes métodos para la evaluación del criterio de selección ó "*fitness*". Métodos basados en redes neuronales de entrenamiento rápido (Fuzzy ARTMAP, PNN) acoplados a los diferentes métodos de selección desarrollados.
- Implementación de métodos de selección de variables basados en Support Vector Machines (SVM) para procesos de reconocimiento y selección de variables en sistemas olfativos.

- Desarrollo de diferentes técnicas de selección de variables de baja carga computacional que permitan eliminar variables ruidosas o con información irrelevante (técnicas de varianza y colinealidad).
- Evaluación de los métodos desarrollados para aplicaciones basadas en sensores de estado sólido y en aplicaciones basadas en espectrometría de masas.

1.3 Organización de la memoria

Este documento consta de 5 capítulos y un anexo con las publicaciones generadas durante el desarrollo del trabajo de investigación que se presenta en esta memoria.

Tras el capítulo de introducción, donde se argumenta sobre el interés científico y técnico de los objetivos de esta tesis, el segundo capítulo presenta tanto los conocimientos fundamentales necesarios para seguir sin dificultad la descripción del desarrollo del trabajo como el estado del arte en el tema de la selección de variables.

En el tercer capítulo, “bases teóricas” se describen con mayor profundidad los conceptos teóricos relacionados con los diferentes métodos de selección de variables desarrollados en este trabajo. En la primera parte se describen las técnicas de reconocimiento de patrones utilizadas, como las redes neuronales fuzzy ARTMAP y PNN o los Support Vector Machines (SVM). Seguidamente se describen uno por uno los diferentes métodos de selección de variables implementados, tanto los secuenciales como los estocásticos. También se describen otras técnicas que no caben en esas definiciones como el método de varianza y el de colinealidad.

En el capítulo cuarto se muestran los diferentes resultados obtenidos al implementar los métodos mencionados anteriormente en cuatro problemas prácticos:

- La identificación de mezclas simples y binarias con vapores de acetona, amoníaco y ortoxileno.
- La clasificación de diferentes disoluciones con impurezas agregadas como el tricloroetieno, el 1- butanol, el etilbenzeno y el tolueno.

- La clasificación e identificación de un conjunto de muestras de aceite de oliva virgen de la región de Tarragona.
- La identificación y clasificación de un conjunto de muestras de jamón ibérico español.

Finalmente, en el capítulo quinto se argumentan las conclusiones obtenidas tras la realización de los estudios descritos en los capítulos anteriores.

Por otra parte, en los anexos de la memoria se puede encontrar todas las publicaciones generadas durante el desarrollo de esta tesis doctoral, tanto los trabajos aceptados en congresos como los artículos enviados a revistas internacionales.

1.4 Referencias

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

[2] R. Marsili, SPME-MS-MVA as an electronic nose for the study of off-flavors in milk, J. Agr. Food Chem. 47 (1999) 648–654.

[3] S. Nakata, Y. Kaneda, H. Nakamura, K. Yoshikawa, Detection and quantification of CO gas based on the dynamic response of a ceramic sensor, Chem. Lett. (1991) 1505–1508.

[4] E. Llobet, R. Ionescu, S. Al-Khalifa, J. Brezmes, X. Vilanova, X. Correig, N. Barsan, J.W. Gardner, Multicomponent gas mixture analysis using a single tin oxide sensor and dynamic pattern recognition, IEEE Sens. J. 1 (2001) 207–213.

[5] N. Paulsson, E. Larson, F. Winquist, Extraction and selection of parameters for evaluation of breath alcohol measurement with an electronic nose, Sens. Actuators A 84 (2000) 187–197.

[6] T. Eklov, P. Martensson, I. Lundstrom, Selection of variables for interpreting multivariable gas sensor data, Anal. Chim. Acta 381 (1999) 221–232.

[7] J. Brezmes, P. Cabre, S. Rojo, E. Llobet, X. Vilanova, X. Correig, Discrimination between different samples of olive using variable selection techniques and modified fuzzy ARTMAP neural networks, in: Proceedings of the Ninth International Symposium on Olfaction and Electronic Nose, ISOEN'02, Rome, Italy, vol. 1, 2002, pp. 188–190.

[8] T. Artursson, M. Holmberg, Wavelet transform of electronic tongue data, Sens. Actuators B 87 (2002) 379–391.

[9] L. Xu, W.-J. Zhang, Comparison of different methods for variable selection, Anal. Chim. Acta 446 (2001) 477–483.

[10] J.M. Sutter, J.H. Kalivas, Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection, Microchem. J. 47 (1993) 60–66.

ISBN:978-84-693-4070-7/DL:T-1167-2010

[11] J. Gardner, P. Bartlett, Electronic Noses: Principles and Applications, Oxford Science Publications, Oxford, 1999.

[12] L. Nolle, D.A. Armstrong, A.A. Hopgood, J.A. Wware, Simulated annealing and genetic algorithms applied to finishing mill optimisation for hot rolling of wide steel strip, Int. J. Knowl.-Based Intell. Eng. Sys. 6 (2002) 104–111.

[13] Vinaixa, M. Llobet, E. Brezmes, J. Vilanova, X. Correig, X " *A fuzzy ARTMAP and PLS based MS e-nose for the qualitative and quantitative assessment of rancidity in crisps*" Sensor and Actuators B, 106 (677 -686), (2005).

[14] S. Rezzi, D. Axelson, K. H´eberger, F. Reniero, C. Mariani, C. Guillou "*Classification of olive oils using high throughput flow H-NMR fingerprinting with principal component analysis, linear discriminant analysis and probabilistic neural networks*" Analytica Chimica Acta 552 (2005) 13–24.

[15] Boronat, M Julia. Esteve, M.Dolores. Aragon, Pilar. "la espectrometría de masas y el aroma del vino" Ediciones y promociones (1999).

[16] Esteban, Luis. "la espectrometría de masas en imágenes" ACK editores (1993).

[17] M. Vinaixa, A. Vergara, C. Duran, E. Llobet, C. Badia, J. Brezmes,X. Vilanova, X. Correig, Fast detection of rancidity in potato crisps using e-noses based on mass spectrometry or gas sensors, Sens Actuators B, in press.

[18] M. Adechy, V.P. Shiers, J.B. Rossell, Study of rancidity and resistance to oxidation in edible oils and fats using electronic nose technology in comparison with conventional techniques, Leatherhead Food RA Research Reports 751, 1998.

[19] R.T. Marsili, SPME-MS-MVA as an electronic nose for the study of off-flavors in milk, J. Agric. Food Chem. 47 (1999) 648–654.

[20] E. Schaller, S. Zenh¨ausern, T. Zesiger, J.O. Bosset, F. Escher, Use of preconcentration techniques applied to a MS-based electronic nose, Analysis 28 (2000) 743–749.

[21] B. Dittmann, S. Nitz, Strategies for the development of reliable QA/QC methods when working with mass spectrometry-based chemosensory systems, Sens. Actuators B 69 (2000) 253–257.

[22] P. Geladi, B.R. Kowalski, Partial least squares regression: a tutorial, Anal. Chim. Acta 185 (1986) 1–17.

[23] R. Leardi, M.B. Seasholttz, R.J. Pell, Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data, Anal. Chim. Acta 461 (2002) 189–200.

[24] D. Broadhurst, 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 (1997) 71–86.

[25] R. Marsili, Flavor, Fragrance and Odor Analysis, Marcel Dekker, New York, 2002.

UNIVERSITAT ROVIRA I VIRGILI
DESARROLLO DE DIFERENTES MÉTODOS DE SELECCIÓN
DE VARIABLES PARA SISTEMAS MULTISENTORIALES
Oscar Eduardo Gualdron Guerrero
ISBN:978-84-693-4070-7/DL:T-1167-2010

2.

Estado del Arte

ISBN:978-84-693-4070-7/DL:T-1167-2010

2.1 Introducción

En este capítulo se presentan los conceptos básicos referentes a los sistemas de olfato artificial, incluyendo tanto los sistemas basados en sensores químicos como los que utilizan técnicas analíticas como la espectrometría de masas, mostrando sus diferentes ventajas así como los principales inconvenientes que pueden tener cada una de las aproximaciones. Adicionalmente, y más en consonancia con el tema tratado en esta tesis, en este capítulo se incluye también un minucioso estudio sobre el estado del arte relacionado con el tema de la selección de variables. Este estudio incluye una recopilación de los principales métodos empleados por otros investigadores para tratar esta problemática tanto en problemas genéricos como en sistemas de olfato electrónico.

2.2 Nociones básicas sobre sistemas de olfato electrónico

La definición más comunmente aceptada de lo que es un sistema de olfato electrónico es la que lo describe como "un instrumento que comprende una matriz de sensores químicos con sensibilidades solapadas y un avanzado sistema de reconocimiento de patrones, capaz de reconocer aromas simples y/o complejos" [1,2].

De una manera coloquial se podría afirmar que a este tipo de instrumentos se les denomina sistemas de olfato electrónico o "narices electrónicas" por dos motivos:

- Porque su configuración y funcionamiento emulan al del sistema de olfato humano.
- Porque pretende realizar funciones tradicionalmente atribuidas al sistema de olfato biológico.

Una de las formas de definir a este tipo de sistemas es diferenciarlos de la instrumentación química tradicional ya que la filosofía de análisis es la que diferencia a ambos tipos de instrumentos. Mientras que en instrumentos tradicionales, como en un cromatógrafo de gases, se caracteriza una muestra

ISBN:978-84-693-4070-7/DL:T-1167-2010

identificando y cuantificando cada componente por separado, los sistemas de olfato electrónico valoran la muestra en su conjunto, sin preocuparse por los componentes individuales que conforman la mezcla gaseosa a caracterizar.

Bajo este concepto, la configuración genérica que responde funcionalmente a la definición anteriormente presentada comprende un sistema de muestreo, un sistema de medición provisto de sensores químicos y un sistema informático que controla el proceso de medición y permite aplicar técnicas de pre-procesado de datos y reconocimiento de patrones para la detección, identificación o cuantificación de cualquier compuesto volátil o aroma.

2.2.1 El sistema de olfato humano

Para entender el funcionamiento de un sistema de olfato electrónico, primero describiremos brevemente como funciona el sistema de olfato biológico (ver figura 2.1).

El sistema olfativo es un potente y complejo sentido que contiene millones de células receptoras, aunque se cree que solo existen entre 300 y 1000 tipos diferentes de encimas quimio-receptoras.

El proceso comienza cuando algunas moléculas de aire o del aroma a detectar entran por el conducto nasal y son capturadas y disueltas en una membrana mucosa en el interior de la glándula olfativa. Cuando son disueltas, las moléculas estimulan la membrana donde se encuentran alojadas las células receptoras o cilia, lo que provoca que las células generen impulsos que van al bulbo olfativo en la región límbica del cerebro. Esta información es enviada de forma simultánea pero con diferentes señales por los receptores olfativos, formando un patrón de señales eléctricas que el cerebro interpreta y reconoce como un aroma característico.

En todo este proceso es importante resaltar la función que realiza el cerebro al recibir estas señales. Se cree que gracias al procesado de los impulsos eléctricos entrantes, el cerebro es capaz de discernir entre unos diez mil aromas diferentes (a pesar de tener solamente entre 300 y 1000 tipos de receptores diferentes) e incrementa la sensibilidad hasta en tres órdenes de magnitud.

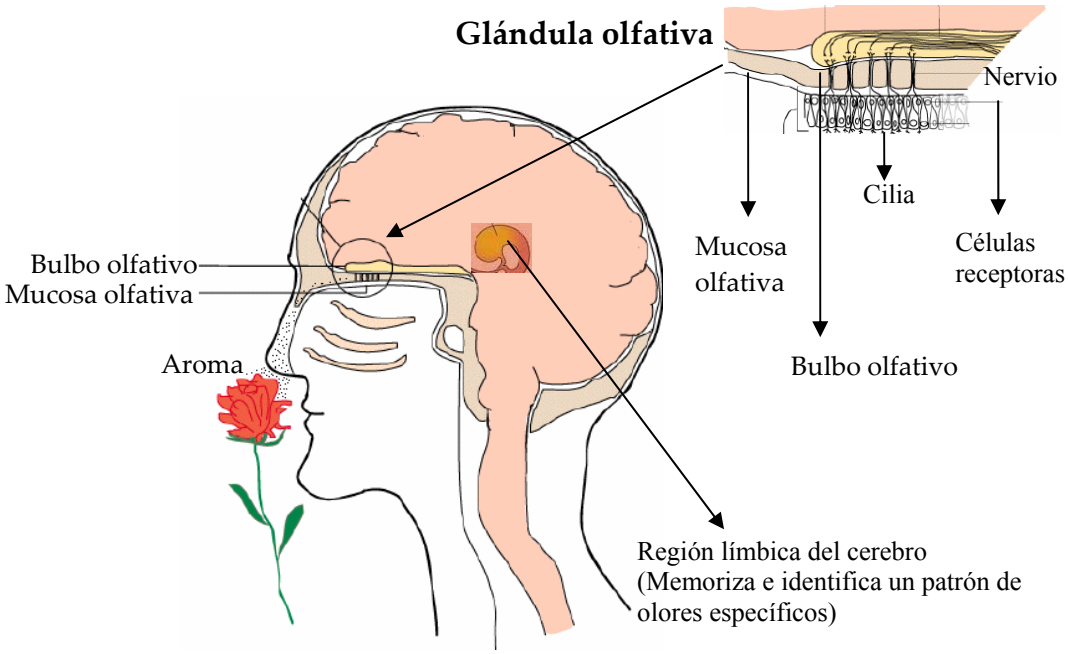

Figura 2.1: Proceso de detección del sistema de olfato humano

2.2.2 Paralelismo con el sistema de olfato artificial

En el esquema de la figura 2.2 se establece un paralelismo entre los componentes que conforman el sistema de olfato biológico frente al artificial. Como podemos observar, los receptores olfativos están representados por un grupo de sensores químicos (matriz de sensores) que producen una señal eléctrica dependiente del tipo de aroma detectado. El bulbo olfativo recibe estas señales para posteriormente enviarlas al cerebro a través de los nervios que, de forma muy similar a las técnicas de preprocesado, preparan las señales reduciendo el volumen de información y minimizando el ruido y las derivas que introducen los receptores olfativos.

ISBN:978-84-693-4070-7/DL:T-1167-2010

Esta tarea facilita la clasificación o identificación de la muestra por medio de un sistema de reconocimiento de patrones (biológico o artificial). Los métodos de reconocimiento de patrones realizan funciones equivalentes a las que se realizan en la corteza del cerebro, que es la etapa final del proceso olfativo humano, donde se identifican, clasifican, interpretan y memorizan los diferentes aromas aprendidos a lo largo del tiempo.

En definitiva, los sistemas de olfato electrónico (SDOE) imitan al sistema olfativo humano, acoplando una matriz de sensores químicos a métodos avanzados de reconocimiento de patrones que permiten caracterizar o discriminar mezclas aromáticas complejas sin una previa separación de sus constituyentes [3,4].

De la misma manera que el olfato biológico no necesita identificar cada compuesto de una mezcla para identificarla, el SDOE valora las muestras aromáticas en su conjunto, sin identificar los componentes básicos que constituyen un aroma complejo.

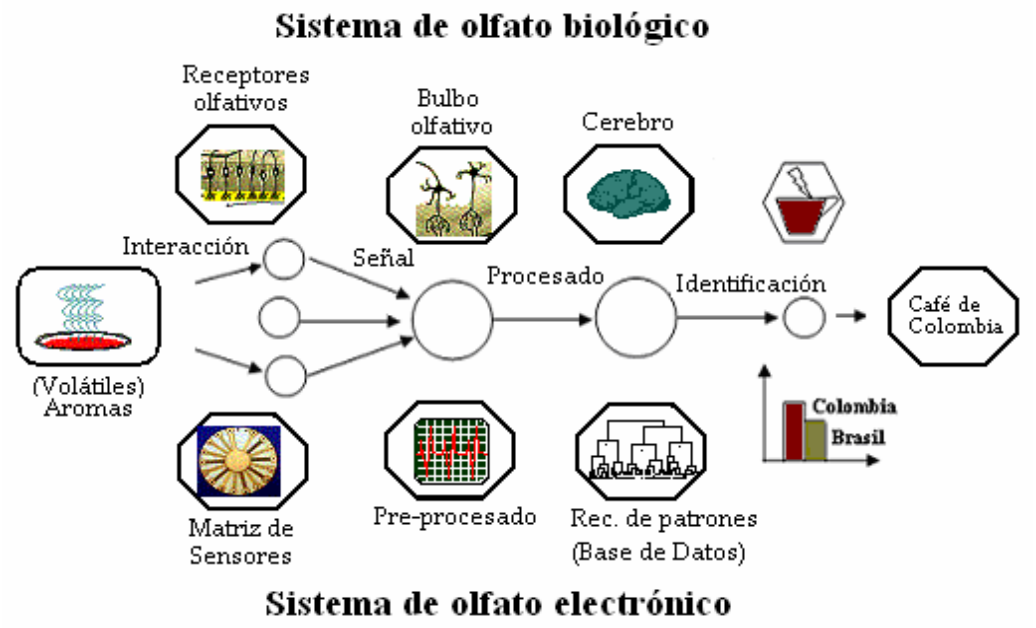

Figura2.2: Estructura de los sistemas de olfato biológico y electrónico

Para entender el funcionamiento de un SDOE es fundamental entender que los sensores químicos que componen la matriz sensorial no son específicos, sino todo lo contrario, son sensores con sensibilidades solapadas. Esto significa que no son selectivos a un compuesto químico dado, pero si levemente más sensibles a determinadas familias químicas, tales como solventes orgánicos, ácidos grasos, gases sulfurosos, etc. De esta forma, las respuestas de los sensores producen señales características para cada mezcla química, siendo sensibles a una amplia variedad de productos químicos.

Las figuras 2.3 y 2.4 muestran este principio mediante un ejemplo ficticio. En la figura 2.3, se puede ver la curva de sensibilidad de tres sensores diferentes ante un espectro de aromas, eje en el que se ha señalado ficticiamente la posición de los aromas de naranja, manzana, uva y pera. La figura 2.4 muestra las respuestas en forma de diagrama radar, de las señales que se obtendrían para cada compuesto. Se puede observar como cada aroma tiene una forma característica que lo identifica. Todo ello gracias a que los sensores presentan sensibilidades solapadas sin ser esencialmente selectivos.

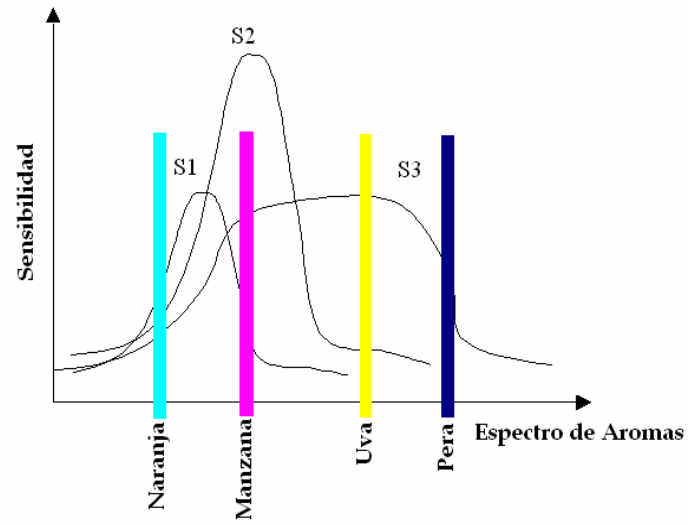

Figura 2.3: Curva de sensibilidades solapadas de tres sensores diferentes, ante un espectro de aromas

ISBN:978-84-693-4070-7/DL:T-1167-2010

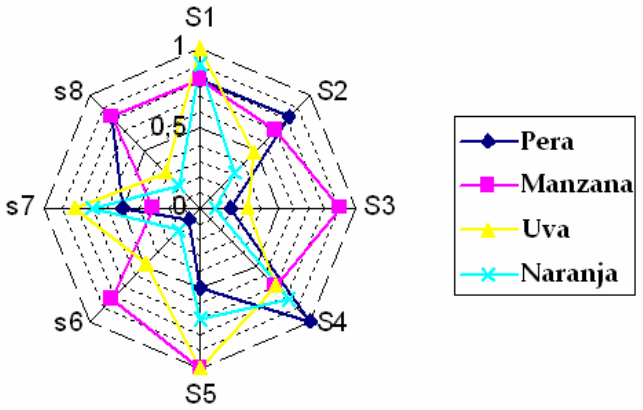

Figura 2.4: Gráfico Radar representando cada compuesto en concentraciones diferentes

A medida que se van realizando nuevas mediciones se va generando una base de datos que se utiliza para entrenar un sistema de reconocimiento de patrones, base de datos que luego permitirá reconocer cada uno de los patrones almacenados en memoria si se vuelve a presentar al sistema.

2.2.3 Módulos básicos y secuencia de trabajo

En la figura 2.5 se observa la secuencia de trabajo y los principales módulos que intervienen en el análisis de aromas mediante un sistema de olfato electrónico. Inicialmente la muestra es acondicionada por métodos de extracción de volátiles que permiten el paso del gas a analizar hacia una matriz de sensores. El sistema de muestreo esta integrado principalmente por un lugar donde se aloja la muestra (como una cámara de concentración), un sistema de control y un sistema de transporte de flujo (como una bomba de aire, controladores de flujo másico, etc).

Existen varias técnicas de manipulación y suministro de flujo hacia la cámara de medida que pueden ser acopladas al sistema en función de la aplicación deseada. Principalmente destacan los sistemas de muestreo por espacio de cabeza estático, dinámico (por ejemplo el sistema de flujo continuo), y técnicas de desorción térmica.

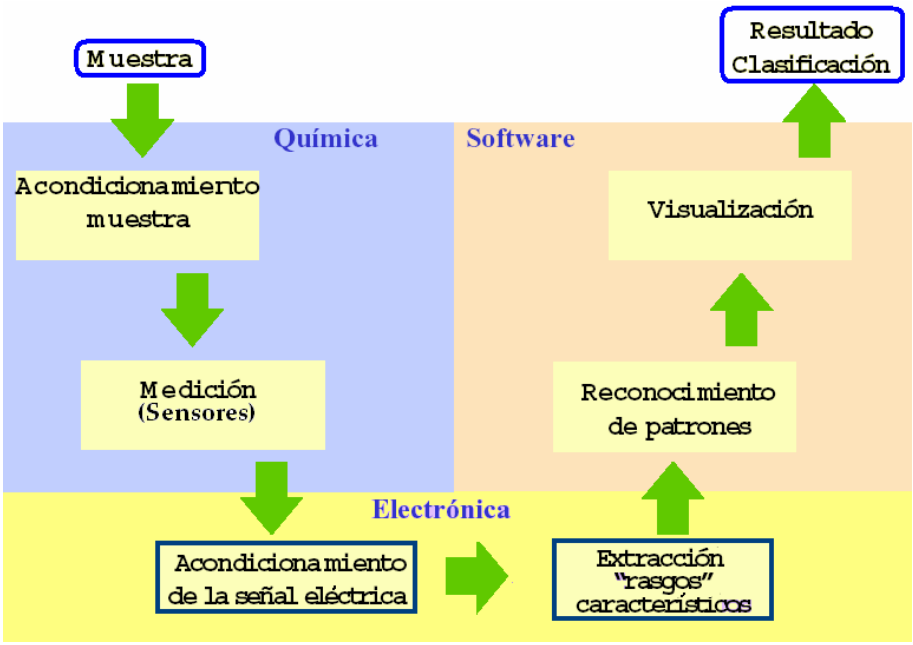

Figura 2.5: Secuencia para el análisis de aromas

El sistema de medición esta compuesto por sensores de gases que cambian sus propiedades físicas en función del entorno gaseoso en el que se vea inmersa la capa activa del dispositivo. Estos cambios se traducen en una respuesta eléctrica generando así una señal dependiente de la presencia de las concentraciones de sustancias que se quieren medir. Esta respuesta generada en el dominio eléctrico es acondicionada para ser leída y almacenada en un ordenador. En el caso de los sistemas de olfato electrónico basados en espectrometría de masas, la respuesta de los sensores es substituida por el espectro de masas que genera dicho instrumento, de forma que la intensidad de cada relación masa/carga (m/z) es considerada un sensor independiente.

Tras ser adquiridas y almacenadas, las señales son tratadas por métodos de extracción de parámetros y pre-procesado de datos. La técnica de extracción de parámetros es fundamental, especialmente al utilizar sensores de oxido de estaño. Estos basan su funcionamiento en el cambio de conductividad que experimenta el

ISBN:978-84-693-4070-7/DL:T-1167-2010

material (capa activa) del sensor ante la presencia de gases reductores y/o oxidantes. El cambio de conductividad experimenta transitorios que llevan a la capa activa del sensor desde una situación de reposo a una conductancia final que depende del tipo de volátil y de su concentración [5]. La información que se puede extraer del sensor es muy variada.

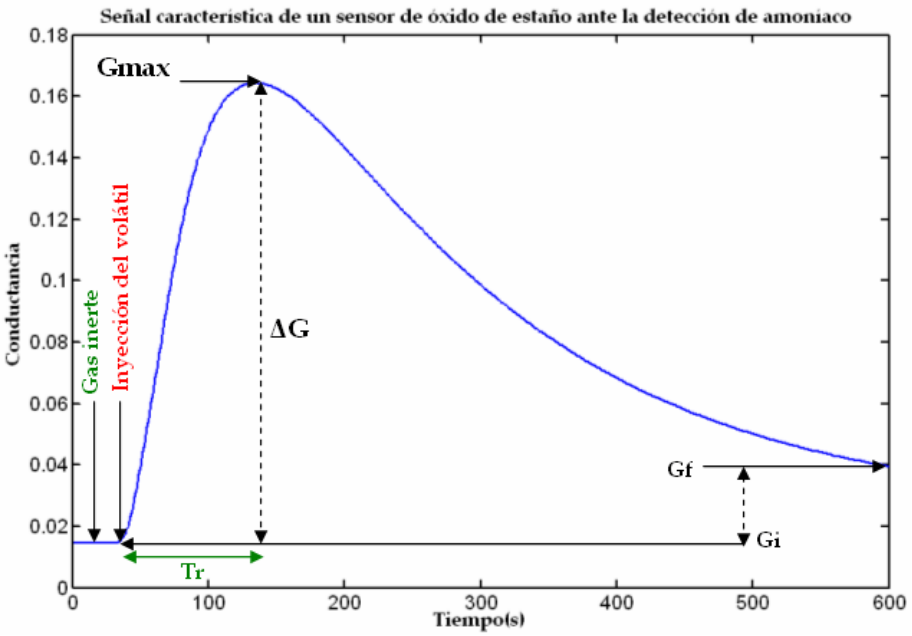

Figura 2.6: Extracción de parámetros temporales ante una respuesta de un sensor de óxido de estaño

La figura 2.6 muestra el transitorio de conductancia típico de un sensor de gas de óxido de estaño frente a un cambio rápido en las concentraciones de amoníaco, donde se puede observar como el transitorio (Tr) tarda unos segundos en llegar a un máximo y como del se pueden extraer parámetros estáticos y parámetros dinámicos.

Si se obtienen valores de conductancia inicial y/o final se dice que se están usando parámetros estáticos. Con ellos se pueden realizar varias combinaciones que son detallas en la siguiente tabla.

Gi	Conductancia inicial
Gf	Conductancia final
Gmax	Conductancia máxima
Δg	Incremento de la conductancia (Gf-Gi) ó (Gmax-Gi)
Δgn	Incremento de la conductancia normalizada (Δg/Gi)

Tabla 2.1: Parámetros estáticos extraídos de las señales de un sensor de óxido de estaño

El objetivo de los métodos de pre-procesado es obtener un vector de datos descriptivo de cada medida que pueda ser procesado por técnicas de reconocimiento de patrones con el fin de analizar y clasificar los compuesto volátiles. Por lo tanto, una vez que los sensores del olfato electrónico reaccionan ante una muestra, se debería procesar los datos obtenidos mediante algoritmos de reconocimiento de patrones que permitan dar la funcionalidad deseada al equipo.

Las diferentes empresas que comercializan equipos de olfato electrónico usan a menudo redes neuronales artificiales (Artificial Neural Networks, ANN) para el reconocimiento de patrones. Los sistemas de redes neuronales tienen muchos elementos de procesos interconectados, como las neuronas en el cerebro. Se puede enseñar a una red a solucionar un problema, tal como reconocer olores y compararlos con los olores que se han analizado y se han almacenado previamente. Cuando se combina una ANN con una matriz de sensores se puede identificar mas olores que número de sensores disponibles, tal y como ocurre en el sistema biológico de olfato.

2.2.4 Ventajas de los sistemas de olfato electrónico

Diversos sectores industriales (como el farmacéutico o el de la alimentación, entre otros) necesitan sistemas fiables para asegurar la calidad y seguridad de su materia prima. En la actualidad, la mayoría de sistemas empleados por los laboratorios de calidad se basan en instrumental analítico tradicional. Éste suele ser muy caro, de difícil puesta a punto, mantenimiento y operación. Además el proceso de análisis

ISBN:978-84-693-4070-7/DL:T-1167-2010

puede ser lento y complicado, lo que implica la contratación de operarios especializados y no suele permitir un análisis en tiempo real. Es en esta situación en la que tiene sentido hablar de sistemas de olfato electrónico. Entre las ventajas que podrían aportar las narices electrónicas se pueden destacar las siguientes:

- Análisis no destructivo del producto.
- Obtención de resultados en tiempo real (en cuestión de minutos)
- Portabilidad, robustez y bajo precio.
- Adaptación a diferentes cantidades y variedades de productos.

2.2.5 Limitaciones actuales de las narices electrónicas

Los sistemas de olfato electrónico parecen tener un gran potencial en la industria alimentaria. Sin embargo, a pesar del gran esfuerzo que se esta dedicando en los laboratorios de investigación, su implantación en la industria es todavía incipiente. Esto puede ser debido a una serie de limitaciones que destacaremos a continuación:

2.2.5.1 Lentitud entre medidas:

Un problema común en las narices electrónicas es determinar el tiempo adecuado de reposo entre medidas. Tras absorber los volátiles al ser expuestos a un flujo de gas, el sensor sigue un proceso de desorpción, que de no completarse, puede afectar a la medida siguiente. A este fenómeno se le conoce como efecto memoria. Hay que destacar que esta limitación no afecta a los sistemas de olfato electrónico basados en espectrometría de masas.

2.2.5.2 Deriva de los sensores:

El objetivo de un sensor químico es dar siempre la misma respuesta cuando es expuesto a muestras idénticas. Sin embargo en la mayoría de sensores que se incorporan en un SDOE esto no es cierto a lo largo de un tiempo prolongado de uso.

Las derivas pueden ser debidas a variaciones de temperatura en el espacio de cabeza, cambios en el sistema de muestreo, envejecimiento de los sensores, variaciones en el flujo de gas, variaciones de humedad y temperatura en la superficie de los sensores, variaciones en la presión ambiental u otros efectos químicos y físicos que influyen en la respuesta del sensor.

La falta de repetitividad en el muestreo es otro de los problemas que afectan a los sistemas de olfato electrónico. Los principales elementos que pueden influir son los siguientes: error experimental, métodos inadecuados en la preparación de las muestras, factores ambientales, etc.

El problema de la deriva de los sensores es mucho menor en los sistemas de espectrometría de masas, aunque también se deben tomar ciertas precauciones para evitar la falta de repetitividad tanto en el muestreo como en los resultados.

2.2.5.3 Baja sensibilidad y selectividad:

El mal acondicionamiento e inapropiado tratamiento de la muestra son la causa fundamental por la que los volátiles no llegan a los sensores en forma óptima. Algunas muestras contienen los volátiles de interés en el rango de las ppb (partes por billón, 10^{-9}) o incluso a concentraciones inferiores y los sensores responden habitualmente en el rango de los ppm (partes por millón, 10^{-6}). Los inadecuados niveles de concentración de las muestras podrían hacer que los sensores no respondan correctamente a diferentes tipos de volátiles.

Por otro lado, a pesar de que el concepto de sensibilidades solapadas es una pieza fundamental del modo de funcionamiento de los SDOE, la falta de selectividad del sensor también es un problema que afecta a estos sistemas debido a que no son capaces de distinguir entre diferentes tipos de muestra. Por lo tanto es fundamental utilizar técnicas que permitan incrementar la selectividad, aunque el precio a pagar sea el incremento de la dimensionalidad del problema a resolver.

Estos dos problemas son compartidos por los sistemas de olfato electrónico basados en espectrometría de masas, por lo que tanto el proceso de muestreo como

ISBN:978-84-693-4070-7/DL:T-1167-2010

de tratamiento de señales (pre y post- procesado) son aspectos fundamentales a perfeccionar si se quieren neutralizar los problemas de sensibilidad y selectividad.

2.2.5.4 Conjunto de entrenamiento elevado

Es importante afirmar que el entreno es algo muy costoso y laborioso de realizar debido a que al sustituir sensores es necesario entrenar nuevamente el sistema. En el momento de la calibración se tienen que pasar muestras de forma periódica lo cual implica una gran carga de trabajo y pérdida de tiempo. El presente proyecto de tesis, en cierta forma, se centra en el estudio de esta limitación buscando la forma de reducir la dimensionalidad del conjunto que se utilizará para el entrenamiento y validación a través de técnicas de selección de variables.

El entrenamiento con un gran número de medidas a través de redes neuronales artificiales sería lo mas conveniente, pero podría llegar a ser un problema debido a que necesitarían encontrar no sólo una configuración óptima, si no también un algoritmo que además de rapidez garantice estabilidad en el resultado final. No existe un procedimiento establecido para determinar qué modelo de red debe emplearse en cada aplicación, y solo con la práctica puede determinarse cual es la configuración de red que da mejores resultados.

2.3 Sistemas de Olfato Electrónico basados en Espectrometría de masas (SM)

Los espectrómetros de masas funcionan habitualmente asociados a un sistema de cromatografía de gases conformando un sistema CG/MS. La etapa cromatográfica (GC) se encarga de separar en el tiempo los diferentes componentes de una muestra aromática compleja (lo que permite obtener picos cromatográficos resueltos a lo largo del tiempo). Dichos picos entran secuencialmente en la etapa de espectrometría de masas, lo que permite identificar a los componentes de la muestra (cada compuesto tiene un patrón de ionización que le es propio).

Por el contrario, los sistemas de nariz electrónica basados en MS eliminan la etapa previa de separación cromatográfica. Por lo tanto, se inyecta al MS una muestra compleja de compuestos volátiles sin haber realizado un paso previo de separación. Como resultado, el espectrómetro de masas obtiene un patrón de ionización complejo correspondiente a la muestra analizada. Todos los compuestos volátiles se ionizan y fragmentan a la vez, lo que puede producir efectos no lineales debido a la interacción entre todos los fragmentos. En cualquier caso, los espectros resultantes son analizados y clasificados por el sistema de reconocimiento de patrones [6].

El espectrómetro de masas usa la diferencia entre los espectros masa-carga (m/z) de las diferentes moléculas presentes en la mezcla gaseosa para identificar o clasificar muestras. Se trata de una técnica útil para la cuantificación de átomos o moléculas y también para la identificación química de las moléculas, así como para suministrar la información estructural sobre las mismas. Las moléculas tienen modelos de fragmentación distintos que permiten identificar los componentes estructurales que las conforman [7]. En el funcionamiento general de un espectrómetro de masas se pueden distinguir tres fases:

- Rotura de las moléculas creando patrones de fragmentación iónicos.
- Separación de los fragmentos iónicos en espacio o tiempo basándose en su relación masa-carga.
- Contado de iones para cada relación masa-carga.

El poder de separación del ion en un espectrómetro de masas es descrito por su resolución, la cual se define como:

$$R = m/\Delta m \qquad (2.1)$$

donde m es la masa del ion y Δm es la diferencia en masa entre dos picos en un espectro de masas. Por ejemplo, un espectrómetro de masa con una resolución de 1000 puede diferenciar un ion con una relación m/z de 100.0 de un ion con un m/z de 100.1.

ISBN:978-84-693-4070-7/DL:T-1167-2010

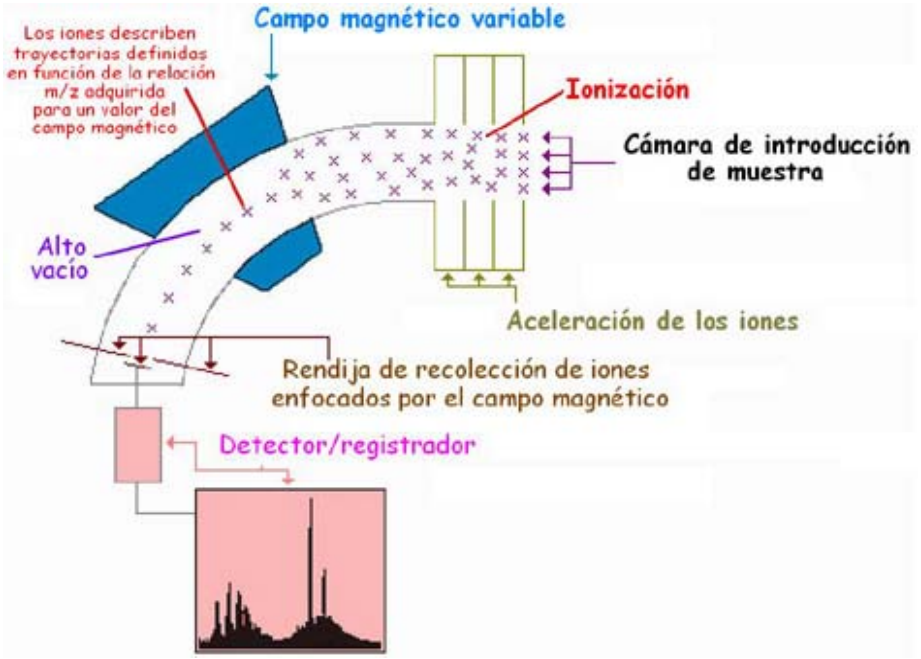

Figura 2.7 Esquema del funcionamiento de un espectrómetro de masas

2.3.1 Partes de un espectrómetro de masas:

Un espectrómetro de masas básicamente está compuesto por una fuente de ionización, un analizador selectivo de masas y un detector de iones [8]. Este proceso se ilustra en la figura 2.7. A continuación definimos más detalladamente cada una de las diferentes partes que conforman el sistema.

2.3.1.1 Entrada

La introducción de la muestra en el interior del espectrómetro se realiza de diferentes maneras, dependiendo de la naturaleza de la muestra. El dispositivo de inyección debe estar diseñado para situar la muestra en el interior el equipo, donde la presión es normalmente inferior a 10^{-6} milibares, y vaporizarla en el caso de que no sea gaseosa. En algunos espectrómetros de masas hay dos zonas con vacíos

diferentes, y en este caso la presión en el punto de inyección de muestra puede ser bastante más alta.

2.3.1.2 Ionización

Una vez la muestra situada ha logrado entrar en el interior del espectrómetro de masas, se procede a su ionización mediante diferentes métodos, según el tipo de muestra que estemos analizando. El sistema de ionización mas usado es el de impacto electrónico o "EI", que bombardea la molécula con electrones de una cierta energía, capaces de provocar la emisión estimulada de un electrón de la molécula, y así ionizarla.

Además de moléculas ionizadas, o iones moleculares, también se forman iones fragmentados debido a la descomposición de iones moleculares con exceso de energía. El tipo y proporción relativa de cada uno de estos fragmentos es característico de la molécula analizada y de las condiciones del proceso de ionización, y se denomina "patrón de fragmentación".

La zona del espectrómetro donde se realiza la entrada y la ionización de la muestra se denomina fuente de ionización o fuente de iones.

2.3.1.3 Aceleración

Una vez que se consigue ionizar las moléculas de la muestra, en el caso de los espectrómetros de masas magnéticos, estos fragmentos se aceleran mediante campos eléctricos que comunican una misma energía cinética a todos los iones formados. La velocidad adquirida por cada ion dependerá de su masa.

2.3.1.4 Análisis

Los iones seguirán una trayectoria forzada mediante campos eléctricos o magnéticos situados en la zona denominada analizador. Sufrirán una mayor o menor desviación, para un mismo valor de la fuerza aplicada, en función de su masa o

ISBN:978-84-693-4070-7/DL:T-1167-2010

velocidad. Variando el valor del campo aplicado entre determinados limites, podemos ir dirigiendo de modo consecutivo los iones de diferentes masas, en orden creciente o decreciente, hacia el sistema colector.

2.3.1.5 Detección

La detección consecutiva de los iones formados a partir de las moléculas de la muestra, suponiendo que se trate de una sustancia pura, produce el espectro de masas de esa sustancia, que es diferente para cada compuesto químico, y que constituye una identificación prácticamente inequívoca del compuesto analizado.

La colección de los iones en el detector (llamado normalmente colector), produce una señal eléctrica que, convenientemente amplificada, es registrada y representada gráficamente a través de una pantalla de ordenador y una impresora. El espectro de masas así obtenido puede almacenarse en la memoria del ordenador, puede compararse con los espectros de una colección de espectros (o espectroteca) para su identificación, puede estudiarse para averiguar la naturaleza de la molécula que le dio origen, etc.

2.3.2 Ventajas de la espectrometría de masas

La espectrometría de masas es una de las técnicas analíticas mas utilizada hoy en día. Entre las principales ventajas se pueden destacar las siguientes:

- Su capacidad de identificación permite determinar cualitativamente y de forma muy precisa casi cualquier tipo de sustancia, desde átomos o compuestos sencillos hasta moléculas extraordinariamente complejas.
- Es cuantitativa y cualitativa. No sólo es capaz de identificar las sustancias analizadas proporcionando un espectro o “huella digital” de la molécula, sino que también puede cuantificar y medir la concentración de las mismas.
- Posee una gran sensibilidad. Puede detectar prácticamente cualquier elemento en concentraciones del orden de los “ppm” (partes por millón).

- Es universal y especifica. Es decir, puede analizar sustancias o mezclas de sustancias sólidas, liquidas o gaseosas, y también es capaz de detectar y separar una sustancia concreta en presencia de una matriz compleja.
- Puede proporcionar información estructural de la molécula analizada, energía de enlaces, información cinética, físico química, quántica, etc.
- Es una técnica muy rápida. Puede medir un espectro en décimas de segundo. Por ello puede utilizarse para monitorización de procesos, suministrando información en tiempo real sobre la composición de una muestra de gases en un reactor, entre otras posibilidades.

2.3.3 Limitaciones de la espectrometría de masas

A pesar de presentar numerosas ventajas, se trata de una técnica con algunas limitaciones:

- En los espectros resultantes puede existir una alta colinealidad entre cada una de las variables o relación masa-carga (m/z). Esto implica que al analizar una mezcla compleja de compuestos volátiles, diferentes compuestos pueden producir patrones de ionización que comparten determinados conjuntos de iones.
- Elevada dimensionalidad del conjunto de datos: Para cada muestra se devuelve un espectro con cientos de relaciones masa/carga, por lo que el conjunto de descriptores o variables para cada muestra es muy grande, dificultando el posterior trabajo de reconocimiento de patrones.
- Problemas de derivas en el detector. El detector debe ser recalibrado periódicamente.
- El proceso de muestreo es más delicado que en SDOE convencionales. Uno de los métodos más utilizados es la toma de muestra mediante microextracción en fase sólida (SPME). En esta técnica los volátiles de la muestra se concentran mediante su adsorción en la superficie de una fibra recubierta con un polímero. Luego se realiza una desorción térmica en el

ISBN:978-84-693-4070-7/DL:T-1167-2010

puerto de entrada del MS. Lo que puede llevar a una desventaja dado que el proceso de desorción debe ser total, puesto que si no lo es, la fibra pierde capacidad de adsorción/concentración de la muestra o, lo que es peor, existe el riesgo de contaminación cruzada entre muestras.

2.4 Estado del arte

La mayoría de trabajos publicados sobre selección de variables están relacionados con la construcción de modelos de regresión cuantitativos (técnicas quimiométricas). Por el contrario, existe un número muy reducido de publicaciones que traten el problema de la selección de variables para sistemas de olfato electrónico. A continuación se revisa brevemente la literatura existente.

K. Tang y colaboradores [9] han trabajado en la utilización de algoritmos genéticos combinados con el método de regresión PLS para implementar la técnica quantitative structure-activity relationships (QSAR). Los métodos utilizados en el estudio QSAR incluyen algoritmos de regresión así como técnicas de reconocimiento de patrones. Sus estudios indican que la combinación del algoritmo PLS con técnicas de selección de variables basadas en algoritmos genéticos (GA) puede ser empleada para describir la relación entre una serie de compuestos y su actividad química. El uso de funciones polinomiales en la relación interna del modelo PLS (modelo PLS no lineal) proporciona un camino directo y simple para modelar las relaciones no lineales existentes ente los datos. Los autores muestran que este método puede ser fácilmente adaptado a cada modelo PLS funcional. Tales modelos pueden, por lo tanto proporcionar un puente entre el modelado empírico y la teoría química fundamental.

Lu Xu y colaboradores [10] estudian muestras multicomponentes mediante quantitative structure-activity/property relationships (QSAR). Sin embargo, a diferencia de K. Tang, no sólo utilizan algoritmos genéticos para la selección de variables, sino que estudian diferentes métodos clásicos como son el forward selection, backward elimination, stepwise regression, orthogonal descriptors y

métodos estocásticos como el leaps and bounds regression (equivalente al branch and bound descrito en [11]). Esto les permite realizar una comparación entre los diversos métodos. Toman como base de su estudio 35 nitrobenzenos con sus correspondientes actividades tóxicas que conforman su conjunto de datos y realizan una selección de variables que les permite correlacionar la presencia de diversos compuestos en diferentes concentraciones con el índice de toxicidad resultante. El conjunto inicial esta formado por un total de 22 variables resultantes de la extracción de los diferentes parámetros físicos relacionados con la estructura molecular de cada uno de los nitrobenzenos y calculados utilizando un equipo MOPAC (molecular orbital package).

En la tabla 2.2 se puede observar que se han tomado como máximo siete variables. La razón es estadística, ya que la proporción del número de muestras (N) al número de variables (m) no debe ser demasiado baja. Usualmente se recomienda que $N/m > 5$. El algoritmo genético como procedimiento de optimización posee la habilidad de investigar un espacio de parámetros más grande y evitar los mínimos locales.

De la tabla se puede deducir que, en varios casos, la variable encontrada como la mejor en el paso previo mediante la técnica stepwise no aparezca como la mejor seleccionada en los pasos siguientes. Por ejemplo observando el algoritmo genético se puede observar que la mejor variable es la 8 en la primera selección, mientras que en la segunda son mejores la 1 y la 6. Lo mismo sucede con el leaps-and-bounds.

De la tabla 2.2 también se puede ver que los resultados obtenidos usando los tres métodos clásicos no son muy buenos comparados con los métodos más novedosos. Esto puede quedar explicado por las limitaciones de los métodos más antiguos. Por ejemplo, en Forward Selection una vez una variable ha entrado en el modelo ya no desaparece.

NUMERO DE VARIABLES	METODO SELECCION	SUBSET MEJORES VARIABLES	COEFICIENTE R	RMS
1	Forward selection	8	0.8329	0.422
	Backward elimination	1	0.82	0.436
	Stepwise regression			
	Algoritmo genético	8	0.8329	0.422
	Leaps and bounds	8	0.8329	0.422
	Orthogonal descript.	1	0.8329	0.422
2	Forward selection	1,8	0.9047	0.324
	Backward elimination	1, 12	0.8490	0.402
	Stepwise regression	1,8	0.9047	0.324
	Algoritmo genético	1,6	0.9070	0.321
	Leaps and bounds	1,6	0.9070	0.321
	Orthogonal descript.	1,2	0.9047	0.324
3	Forward selection	1,2,8	0.9098	0.316
	Backward elimination	1,10,12	0.8895	0.348
	Stepwise regression	1,2,8	0.9098	0.316
	Algoritmo genético	1,6,21	0.9150	0.307
	Leaps and bounds	1,6,21	0.915	0.307
	Orthogonal descript.	1,2,7	0.9278	0.284
4	Forward selection	1,2,6,8	0.9126	0.312
	Backward elimination	1,10,11,12	0.9164	0.305
	Stepwise regression	1,2,6,21	0.9156	0.306
	Algoritmo genético	1,10,11,17	0.9175	0.303
	Leaps and bounds	1,10,11,17	0,9175	0.303
	Orthogonal descript.	1,2,16,17	0.9373	0.265

Tabla2.2.a. Comparación entre los diferentes métodos de selección. El coeficiente R es igual al coeficiente de correlación y RMS indica el error cuadrático medio.

NUMERO DE VARIABLES	METODO SELECCION	SUBSET MEJORES VARIABLES	COEFICIENTE R	RMS
5	Forward selection	1,2,6,8,21	0.9156	0.306
	Backward elimination	1,10,11,12,16	0.9188	0.301
	Stepwise regression			
	Algoritmo genético	1,5,12,17,19	0.9213	0.296
	Leaps and bounds	1,5,12,17,19	0.9213	0.296
	Orthogonal descript.	1,2,10,16,17	0.9456	0.248
6	Forward selection	1,2,6,7,8,21	0.9172	0.303
	Backward elimination	1,10,11,12,16,20	0.9219	0.295
	Stepwise regression	1,2,6,7,8,21	0.9172	0.303
	Algoritmo genético	1,9,10,11,16,17	0.9279	0.284
	Leaps and bounds	1,9,10,11,16,17	0.9279	0.284
	Orthogonal descript.	1,2,10,16,17,18	0.9470	0.245
7	Forward selection	1,2,6,7,8,17,21	0.9188	0.301
	Backward elimination	1,10,11,12,16,20,21	0.9243	0.291
	Stepwise regression	N/D	N/D	N/D
	Algoritmo genético	1,7,9,10,11,12,16	0.9346	0.271
	Leaps and bounds	1,7,9,10,11,12,16	0.9346	0.271
	Orthogonal descript.	1,2,10,15,16,17,18	0.9594	0.215

Tabla2.2.b (Continuación). Comparación entre los diferentes métodos de selección. El coeficiente R es igual al coeficiente de correlación y RMS indica el error cuadrático medio.

Debido a esto, uno nunca estará seguro del nivel de optimización del subconjunto de variables. En el caso del backward elimination el problema radica en que una vez una variable se ha descartado, no puede ser incluida de nuevo, lo cual afecta a la fiabilidad del subconjunto escogido como combinación óptima. Como previamente se mencionó, el método stepwise es esencialmente un forward selection, aunque la variable seleccionada puede quitarse de nuevo. Estos factores pueden llevar a unos resultados como los obtenidos en los que se han encontrado soluciones sub-óptimas.

ISBN:978-84-693-4070-7/DL:T-1167-2010

N. Paulsson y colaboradores en [12] implementan una nariz electrónica para la detección y evaluación de medidas de alcohol en la respiración (prueba de alcoholemia), realizando una extracción y selección de parámetros a través de métodos clásicos como el forward selection combinado con redes neuronales artificiales para predecir las concentraciones de etanol. El conjunto de medidas es de 140, obtenidas de 14 sensores MOSFET (químico-resistivos), con 5 parámetros por sensor y 2 tipos diferentes de normalizaciones para los datos.

En este trabajo, se utiliza el forward selection para la selección de parámetros. Antes de la selección, el conjunto se divide al azar en un conjunto de entrenamiento de 100 medidas y en un conjunto de validación de 40 medidas. Estos conjuntos son también utilizados en la secuencia de evaluación de los datos con la red neuronal. El criterio de fitness para la forward selection es el error de predicción del conjunto de validación y se calcula por medio de la raíz cuadrada del error cuadrático medio (RMSE) de una regresión múltiple lineal basada en la relación entre Z e Y:

$$Y = a + \sum b_i Z_i + e \qquad (2.2)$$

Z es el subconjunto buscado de parámetros, a y b_i son los coeficientes de estimación, Y es la salida (por ejemplo, la concentración de etanol) y e es un residuo. Y y Z son determinados usando el algoritmo de mínimos cuadrados. El RMSE se calcula mediante la expresión:

$$RMSE = \left(\sum (X_p - X_m)^2 / (n-1)\right)^{1/2} \qquad (2.3)$$

dónde X_p es la salida predicha, X_m es la salida medida y n es el número de parámetros.

De los 140 parámetros iniciales se han seleccionado 30 por medio del método de selección aplicado, resultando un error cuadrático medio de 25.7 ppm. En la figura 2.8 se observa el gráfico del error RMSE durante la forward selection mostrando el número de parámetros seleccionados, los cuales son los escogidos para usarlos en el modelo ANN.

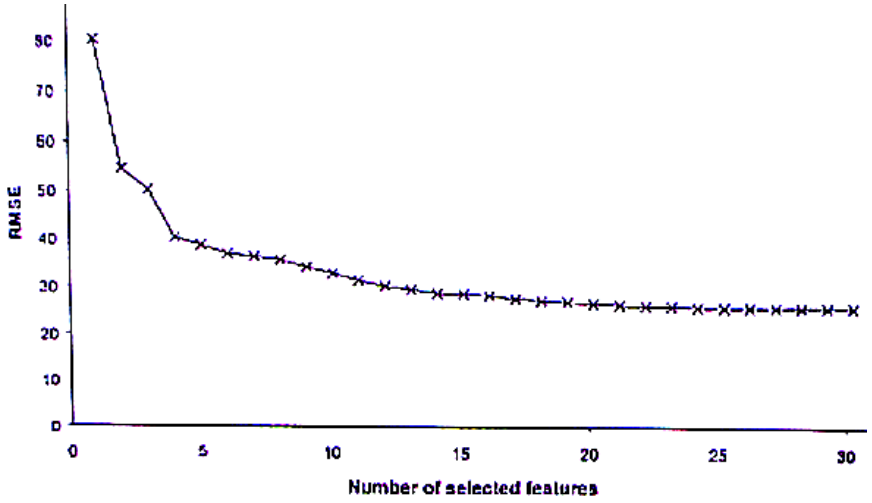

Figura 2.8 gráfico del error de permanencia RMSE durante la forward selection.

T. Eklov y colaboradores examinan en [13] diferentes métodos para seleccionar variables relevantes de un conjunto de variables resultantes de emplear una matriz de sensores de gases. El objetivo es encontrar el mejor subconjunto de parámetros que pueda estimar propiedades interesantes de las medidas. Utilizan forward selection aplicando el error cuadrático medio (RMSE) de un modelo de regresión multilineal como criterio de selección. Con él prueban si el nuevo conjunto obtiene buenas predicciones en una red neuronal de tipo backpropagation. A su vez examinan el uso de vectores de scores (variables secundarias) obtenidos mediante análisis PCA y PLS como método de selección. Utilizan dos conjuntos de medidas, siendo el primero la información extraída de la curva de respuesta de un sensor. Las medidas fueron hechas con un sensor Pt-MOSFET expuesto a diferentes concentraciones (entre 0 y 50 ppm) de hidrogeno y etanol. El segundo conjunto utilizado proviene de un proceso de cultivo bacteriano. Una matriz de multisensores químicos se usó para monitorizar una muestra de E. Coli, con el objetivo de estimar la biomasa y el rango de crecimiento específico. Los resultados de predicción obtenidos fueron mejores comparados con estudios anteriores, empleando los diferentes métodos para reducir la dimensionalidad del conjunto de datos inicial.

ISBN:978-84-693-4070-7/DL:T-1167-2010

J. Brezmes y colaboradores utilizaron en [14] una nariz electrónica para clasificar muestras de aceite de oliva. Emplearon algoritmos de procesamiento de datos incluyendo PCA y redes modificadas fuzzy ARTMAP. Para reducir el número de variables a utilizar y seleccionar un subconjunto óptimo definen dos criterios diferentes. El primer criterio se define como el poder de resolución de cada variable (relación entre la varianza externa y la varianza interna). El segundo criterio es el promedio de sensibilidad para cada sensor. El conjunto de medidas fue de 90 (9 repeticiones para 10 muestras) con un 62 % de tasa de éxito en la clasificación con todas las variables y un 78 % de éxito empleando el proceso de selección.

T. Artusson y colaboradores muestran en [15] que para un particular sistema de lengua electrónica, la compresión de los datos puede hacerse usando transformadas wavelet junto con diferentes algoritmos de selección. El resultado de la compresión de los datos puede también usarse para facilitar la interpretación de las medidas. Utilizaron dos criterios para la selección de los coeficientes wavelet en dos diferentes conjuntos de datos. En el primer conjunto de datos (formado por medidas provenientes de una planta de producción de agua) se utilizó el análisis de componentes principales (PCA) como criterio de selección de los coeficientes. El objetivo de las medidas fue monitorizar el agua después de su tratamiento con diferentes filtros. En este caso, el número de variables se redujo en un factor de 18, sin perder información relevante. El segundo conjunto se centró en la separación de diferentes microorganismos utilizando como criterio de selección de los coeficientes la relación entre las varianzas de una misma clase con respecto a las varianzas de diferentes clases (relación inter/intra varianza), reduciendo el número de variables en un factor de 144. El conjunto de datos reducido capturó suficiente información importante para la identificación de los microorganismos.

En otro artículo publicado recientemente por J.W. Gardner y colaboradores [17], se emplean conocidas técnicas de selección de variables como son los algoritmos genéticos (GA), el forward y backward selection, para encontrar el subconjunto óptimo de sensores dentro de una agrupación. El conjunto de datos está formado por 180 muestras de cultivos de 6 tipos de bacterias responsables de infecciones oculares. Los volátiles emitidos por los cultivos son medidos usando una nariz electrónica con

32 sensores. Se obtiene un 89 % de clasificación con solo 3 variables (3 sensores) empleando el método forward selection, mientras que con la técnica del backward selection necesita como mínimo 5 variables para alcanzar el mismo resultado. Con GA la dimensionalidad se reduce en un 50-60% con un 91 % en el resultado de clasificación usando ocho, seis o cuatro variables, resultados muy similares al obtenido con la totalidad de las variables que fue de un 92 %. En este caso se utilizó una red PNN para el proceso de clasificación.

A. Alexandridis y colaboradores presentan en [18] una novedosa metodología para seleccionar variables en modelos no lineales, combinando las ventajas de varias tecnologías de inteligencia artificial, fundamentalmente la red neuronal RBF (Radial Basis function). Las variables apropiadas son seleccionadas en dos bloques usando una técnica de optimización multi objetivo. En el primer bloque, un diseño especial de algoritmo genético minimiza el error de predicción supervisando el conjunto de datos, mientras que en el segundo bloque se emplea la técnica Simulated Annealing con el fin de reducir el número de variables iniciales. La eficiencia del método propuesto se demostró a través de diferentes conjuntos de datos referenciados en otros trabajos [19].

R. Meiri y colaboradores [20] emplean el método estocástico Simulated Annealing para realizar una selección de variables y comparar los resultados con los obtenidos por modelos de búsqueda más comunes como el Stepwise Regression (SWR). Estos algoritmos fueron aplicados a conjuntos provenientes de base de datos de marketing. Los autores muestran que el algoritmo SA obtiene resultados un poco mejores con respecto al algoritmo SWR, donde la mayor diferencia entre los dos modelos aparece en la estabilidad del algoritmo, siendo SA mucho más estable y casi insensible a variaciones en la optimización de los parámetros. Por contra, el SWR puede estar afectado en mayor grado por fluctuaciones en la optimización de los correspondientes parámetros, siendo necesario probar con varias configuraciones hasta obtener el mejor modelo. A su favor tiene la facilidad de implementación y el tiempo de ejecución, ya que la carga computacional es menor.

ISBN:978-84-693-4070-7/DL:T-1167-2010

Por otro lado, a continuación se comentan algunos trabajos relacionados con la selección de variables en aplicaciones de narices electrónicas basados en espectrometría.

M. Vinaixa y colaboradores [6] introducen un nuevo método para evaluar rancidez y oxidación en patatas usando un sistema de olfato electrónico basado en espectrometría de masas y toma de muestra por SPME. Este método puede representar una alternativa viable comparada con las técnicas tradicionales como son los test ADV y Rancimat, ya que la preparación de las muestras y el análisis son mucho mas rápidos (por ejemplo empleando el método Rancimat puede requerir varias horas para producir un resultado si la muestra analizada es de buena calidad). La efectividad de la nariz electrónica en la evaluación de la calidad de las patatas se demostró desarrollando dos aplicaciones diferentes. Inicialmente la nariz electrónica se usó para clasificar muestras de patatas de acuerdo a su rancidez. El conjunto se procesó utilizando un clasificador fuzzy ARTMAP con un porcentaje de acierto en la clasificación estimado en un 93 % aproximadamente (resultado de la validación). El sistema pudo discriminar un 100 % de las patatas frescas (clase A) con respecto a las rancias (clase B, C y D). La nariz electrónica fue entrenada para predecir los resultados de los test ADV y Rancimat construyendo modelos cuantitativos PLS. Obteniendo una buena correlación entre el sistema empleado y los resultados de los test ADV y Rancimat (los coeficientes de correlación fueron 0.98 y 0.97 respectivamente). Pero el mejor resultado se obtuvo reduciendo la dimensionalidad del conjunto de entrada aplicando procesos de selección de variables basados en análisis de componentes principales (PCA) y algoritmos genéticos.

S. Rezzi y colaboradores [21], presentan como objetivo principal de su trabajo ilustrar la relevancia de la huella dactilar (espectro) suministrada por un espectrómetro de resonancia magnética nuclear (H-NMR) para evaluar el origen geográfico y el año de producción de diferentes aceites de oliva en varias regiones mediterráneas, combinando el sistema H-NMR con técnicas multivariantes. Realizaron un análisis de componentes principales (PCA) sobre un conjunto de aproximadamente 12,000 variables (derivas químicas), definiendo cuatro conjuntos a priori para el PCA. Aplicaron también un análisis discriminante lineal (LDA) a los

50 primeros PC's para poder clasificar las muestras de oliva de acuerdo a su origen geográfico y al año de producción. La correspondiente selección de variables empleando LDA se consiguió usando las cinco mejores variables y un modelo interactivo forward-stepwise selection. Usando LDA sobre el conjunto de validación externo, la clasificación correcta varia entre 47 y 75 % (selección aleatoria) y entre 35 y 92 % (empleando selección Stone-Kennard (KS)) dependiendo del origen geográfico (país) y los años en que se produjeron. Mencionan también que el porcentaje de acierto puede mejorarse significativamente empleando la red neuronal probabilística (PNN), con resultados entre 58 y 100 % en la clasificación sobre los conjuntos de validación

A continuación se describen algunos de los principales trabajos donde se emplean Support Vectors Machines (SVM) en aplicaciones de sistemas olfativos:

Muchos de los estudios publicados intentan clasificar mezclas simples o binarias de vapores comunes. La mayoría de estos trabajos comparan el funcionamiento de los SVM contra paradigmas más tradicionales como las redes neuronales feed forward o la función básica radial (RBF).Por ejemplo, en [22], M.Distante y colaboradores evalúan el funcionamiento de un modelo SVM en una nariz electrónica basada en sensores de gases de óxido de estaño dopados sobre sol-gel. En el estudio analizan siete tipos diferentes de muestras (agua. Acetona, hexanal, pentanone y mezclas binarias entre las ultimas tres) y usan las señales de los sensores para clasificar las muestras, comparando el funcionamiento de los SVM's con respecto a otros métodos.

El modelo SVM construye siete máquinas diferentes para diferenciar cada especie de las restantes. La red se validó usando un leave one out, método utilizado también para determinar el mejor parámetro de regulación (C). Como el problema es linealmente no separable, se empleó una función kernel polinomial de segundo grado para transformar el problema no lineal en un problema linealmente separable.

ISBN:978-84-693-4070-7/DL:T-1167-2010

	Agua	Acetona	M1	Hexanal	M2	M3	Pentanone
Agua	28	0	0	0	0	0	0
Acetona	0	28	0	0	0	0	0
M1	0	0	33	0	3	0	0
Hexanal	0	0	0	34	0	0	1
M2	1	0	4	1	32	0	0
M3	0	0	0	0	1	50	0
Pentanone	0	0	0	0	0	0	24

Tabla 2.3. Matriz de confusión usando SVM

	Agua	Acetona	M1	Hexanal	M2	M3	Pentanone
Agua	19	0	3	0	6	0	0
Acetone	0	26	0	1	0	0	1
Ml	0	0	27	1	8	0	0
Hexanal	0	0	0	33	0	0	2
M2	0	0	7	0	31	0	0
M3	0	0	4	0	3	44	0
Pentanone	0	0	0	1	0	0	23

Tabla 2.4 Matriz de confusión usando RBF

Las tablas 2.3 y 2.4 comparan las matrices de confusión obtenidas por los modelos SVM y RBF respectivamente, donde las filas muestran las clases verdaderas y las columnas muestran las clases estimadas. Puede verse que mientras los SVM tienen un número pequeño de concentraciones erróneas en las mezclas, la red RBF muestra más errores de predicción, algunos de ellos incluso en vapores simples.

Pardo y Sberveglieri miden en [23] diferentes mezclas de cafés usando una nariz electrónica. Esta nariz esta compuesta por cinco sensores de gases semiconductores de capa delgada. El objetivo del estudio fue de evaluar la habilidad de generalización de los SVM's con dos diferentes funciones kernel (polynomial y gausiana) y sus

correspondientes valores kernel. La conforman un total de 36 medidas por cada una de las 7 diferentes mezclas de café analizadas. Para encajar el problema de forma binaria, convierten artificialmente las siete clases en un conjunto de medidas de dos categorías basadas en proyecciones PCA. En este estudio, el parámetro de regulación se fijo en un valor estándar de uno, usando 4 bloques de validación. En el estudio se evalúa el funcionamiento de cada red contra dos parámetros: el número de componente principales retenido de la proyección PCA y un valor kernel (el orden del polinomio en la función polinomial y el valor para el kernel Gausiano). En este estudio se muestra que para SVM's con kernel RBF, el mínimo error se encuentra para valores pequeños de varianza (valores altos dan como resultado un overfiting) y que más de dos componentes PC tengan que ser usados para evitar un bajo fitting. En este caso se determinó un valor de segundo grado para la función kernel polinomial.

S. Al khalifa y colaboradores analizan en [24] niveles de monóxido de carbono y dióxido de nitrógeno encontrados en el aire utilizando dos diferentes sensores de gases resistivos de capa delgada. Dichos sensores tienen la desventaja de presentar un gran consumo de potencia, siendo un inconveniente para analizadores multi-gas portátiles. Los autores reportan también el uso de un sensor de gas resistivo de baja potencia modulado térmicamente para analizar ambos gases. El substrato micromecanizado no solo reduce la potencia de consumo en DC a 100 mW trabajando a 300 ºC, sino que también permite modular la temperatura en AC. Los autores emplean SVM's junto a coeficientes wavelet de la señal en AC (ver figura 2.9). Este método permitió la rápida clasificación de los gases mezclados CO/NO_2 con un alto nivel de fiabilidad (94 % o mas) usando un solo microsensor de gas de baja potencia.

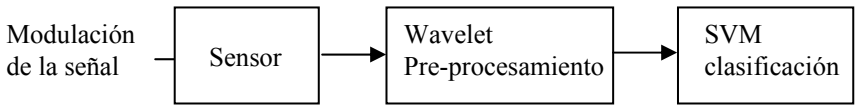

Figura 2.9 Diagrama de bloque del sistema para la detección de monóxido de carbono y dióxido de nitrógeno.

ISBN:978-84-693-4070-7/DL:T-1167-2010

2.5 Conclusiones

Este capítulo comienza con una breve introducción, donde se mencionan las nociones básicas relacionadas con los sistemas de olfato electrónico (sección 2.1), realizando una comparación entre el sistema biológico del olfato humano con respecto a los sistemas de olfato artificial, Incluyendo también la secuencia de trabajo y los principales módulos que intervienen en el análisis de aromas mediante un sistema de olfato electrónico. Además se muestran las diferentes ventajas y las principales limitaciones que se pueden encontrar en estos sistemas, destacando principalmente la limitación de tener conjuntos de entrenamiento elevado, ya que el presente proyecto de tesis, en cierta forma, se centra en el estudio de esta limitación buscando la forma de reducir la dimensionalidad del conjunto que se utilizará para el entrenamiento y validación a través de técnicas de selección de variables.

Por otro lado, en la sección 2.2 de este capítulo también se define el funcionamiento de los sistemas de olfato electrónico basados en espectrometría de masas (MS). De igual forma que en los sistemas de olfato electrónicos se han detallado las ventajas así como las posibles limitaciones que pueden presentar estos sistemas.

Finalmente es importante mencionar que en la sección 2.3 se incluye un minucioso estudio sobre el estado del arte relacionado con el tema de la selección de variables, donde se revisa la literatura existente de los principales métodos empleados por otros investigadores para tratar esta problemática tanto en problemas genéricos como en sistemas de olfato electrónico. Cabe resaltar que la mayoría de trabajos publicados sobre selección de variables están relacionados con técnicas quimiométricas. Por el contrario, existe un número muy reducido de publicaciones que traten el problema de la selección de variables para sistemas de olfato electrónico.

2.6 Referencias

[1] J. W. Gardner and P. Bartlett "A brief history of electronic noses", Sensors and Actuators B, 18-19. 211-220, (1994).

[2] G.H. Dodd, P.N. Bartlett, and J.W. Gardner "Odours--the stimulus for an electronic nose, in Sensors and Sensory Systems for an Electronic Nose" (J.W. Gardner and P.N. Bartlett, Eds.). Proc. NATO Advanced Research Workshop, Reykjavik, Iceland, August 5-8, (1991).

[3] Bartlett, P. N., Elliott, J. M. & Gardner, J. W, "Electronic noses and their applications in the food industry". Food Technology, 51(12), pág: 44-48, (1997).

[4] Haugen, J.-E. & Kvaal, K, "Electronic nose and artificial neural network". Meat Science, 49 (Suppl. 1), pág: 273-286, (1998).

[5] J. Brezmes, X.Correig. Diseño de una nariz electrónica para la determinación no destructiva del grado de maduración de la fruta. Universidad Politécnica de Cataluña, (2001).

[6] Vinaixa, M. Llobet, E. Brezmes, J. Vilanova, X. Correig, X " A fuzzy ARTMAP and PLS based MS e-nose for the qualitative and quantitative assessment of rancidity in crisps" Sensor and Actuators B, 106 (677 -686), (2005).

[7] Boronat, M Julia. Esteve, M.Dolores. Aragon, Pilar. "la espectrometría de masas y el aroma del vino" Ediciones y promociones (1999).

[8] Esteban, Luis. "la espectrometría de masas en imágenes" ACK editores (1993).

[9] Kailing Tang, Tonghua Li "Combining PLS with GA-PLS for QSAR" Chemometrics and intelligent laboratory systems, 64 (2002) 55-64.

[10] Lu Xu, Wen-Jun Zhang, "Comparison of different methods for variable selection", Analytica Chimica Acta 446 (2001) 477-483.

[11] G.M Furnival R. W. Wilson, Technometrics 16 (1974) 499.

ISBN:978-84-693-4070-7/DL:T-1167-2010

[12] Nils Paulsson, Larson Elisabeth, Winquist Fredik "Extraction and selection of parameters for evaluation of breath alcohol measurement with an electronic nose", Sensors Actuators A 84 (2000) 187-197.

[13] Tomas Eklov, per Materson, Ingeman Lundstrom "Selection of variables for interpreting multivariable gas sensor data", Analytica chimica acta 381 (221-232) (1999).

[14] Brezmes, J.; Cabre, P.; Rojo, S.; Llobet, E.; Vilanova, X.; Correig, X., "Discrimination between different samples of olive oil using variable selection techniques and modified fuzzy artmap neural networks", Sensors Journal, IEEE, Volume 5, Issue 3, June 2005 Page(s):463 - 470

[15] Tom Artursson, Martin Holmberg "Wavelet transform of electronic tongue data" Sensors and Actuators B 87 (2002) 379-391.

[16] E. Llobet, 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 Actuators B 99 (2004) 267-272.

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

[18] Alex, Alexandridis. Panagiotis, Patrinos. Haralambos, Sarimveis. George, Tsekouras. "A two-stage evolutionary algorithm for variable selection in the development of RBF neural network models" Chemometrics and Intelligent Laboratory Systems 75 (2005) 149– 162.

[19] J.P. Gauchi, P. Chagnon, "Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data" Chemometr. Intell. Lab. Syst. 58 (2001) 171– 193.

[20] R. Meiri. Jacob, Zahav. "Using simulated annealing to optimize the feature selection problem in marketing applications" European Journal of Operational Research 171 (2006) 842–858.

[21] S. Rezzi, D. Axelson, K. H´eberger, F. Reniero, C. Mariani, C. Guillou "Classification of olive oils using high throughput flow H-NMR fingerprinting with principal component analysis, linear discriminant analysis and probabilistic neural networks" Analytica Chimica Acta 552 (2005) 13–24.

[22] C. Distante, N. Ancona, P. Siciliano. "Support vector Machines for olfactory signals recognition" Sensors and Actuators B, 88, 30-39 (2003).

[23] M. Pardo. G. Sberveglieri. "Classification of electronic nose data with Support Vector Machines" Sensors and Actuators B 107 (2005) 730-737.

[24] Al-Khalifa S., Maldonado-Bascon S., Gardner J.W., Identification of CO and NO_2 using a thermally resistive microsensor and support vector machine, IEE Proceedings measurement and technology, 150, 11-14 (2003).

UNIVERSITAT ROVIRA I VIRGILI
DESARROLLO DE DIFERENTES MÉTODOS DE SELECCIÓN
DE VARIABLES PARA SISTEMAS MULTISENTORIALES
Oscar Eduardo Gualdron Guerrero
ISBN:978-84-693-4070-7/DL:T-1167-2010

UNIVERSITAT ROVIRA I VIRGILI
DESARROLLO DE DIFERENTES MÉTODOS DE SELECCIÓN
DE VARIABLES PARA SISTEMAS MULTISENTORIALES
Oscar Eduardo Gualdron Guerrero
ISBN:978-84-693-4070-7/DL:T-1167-2010

3.

Base teórica y Métodos

ISBN:978-84-693-4070-7/DL:T-1167-2010

3.1 Introducción

En este capítulo se describen con mayor profundidad los conceptos teóricos relacionados con los diferentes métodos de selección de variables desarrollados en este trabajo. En la primera parte se detallan las técnicas de reconocimiento de patrones utilizadas, como las redes neuronales fuzzy ARTMAP y PNN o los Support Vector Machines (SVMs). Seguidamente se describen uno por uno los diferentes métodos de selección de variables implementados, tanto los secuenciales como los estocásticos. Finalmente, también se detallan otras técnicas que no caben en esas definiciones como el método de varianza y el de colinealidad.

3.2 Algoritmos de reconocimientos de patrones

Una de las partes importantes en las NE son el conjunto de técnicas que se utilizan para procesar los datos obtenidos a través de la etapa de sensado. Tales técnicas se denominan algoritmos de "reconocimiento de patrones" o "técnicas de inteligencia artificial" y se pueden definir como los procesos matemáticos que retransforman los datos originales de las medidas en información útil para el usuario final. Actualmente hay un gran número de técnicas de reconocimiento de patrones (PARC) disponibles [1,2,3]. Para seleccionar los algoritmos PARC apropiados en aplicaciones con NE es importante saber la naturaleza fundamental de los datos a analizar.

El problema principal en el análisis de los conjuntos de datos obtenidos por una NE es determinar las relaciones subyacentes entre un conjunto de variables de entrada independientes (como por ejemplo las salidas de una matriz de *n* sensores) y otro conjunto de variables de salida dependientes (como categorizaciones o concentraciones) usando, por ejemplo, un análisis multivariante. En un sistema de análisis de olores generalmente se suele incluir un análisis multivariante y/o métodos PARC que serian usados para analizar cualitativamente los patrones de los olores producidos por esos instrumentos, aunque también podrían ser usados cuantitativamente para calcular concentraciones individuales de los componentes que

ISBN:978-84-693-4070-7/DL:T-1167-2010

componen la muestra analizada. Considerando que el procesado de los datos y el análisis de patrones construyen modelos que relacionan las respuestas de los sensores con resultados interpretables por el usuario, no es de extrañar que este aspecto de los sistemas de olfato electrónico sea un componente fundamental en la implementación, desarrollo y futura comercialización de sistemas multisensoriales (NE).

Estos algoritmos matemáticos pueden ser clasificados de diferentes formas atendiendo a sus características. Así podemos distinguir entre algoritmos supervisados y no supervisados si nos fijamos en el proceso de aprendizaje, algoritmos de clasificación o de cuantificación en función de la naturaleza (binaria o analógica) de la respuesta que deben proporcionar a cada estímulo de entrada, lineales o no lineales según las operaciones que realicen, paramétricos o no paramétricos en función de si se hacen suposiciones iniciales sobre el proceso a modelar, etc:

Métodos paramétricos: Las técnicas paramétricas comúnmente se relacionan a una búsqueda estadística basada en la suposición de que la totalidad de los datos de entrada pueden ser descritos por una función densidad probabilística (PDF). En muchos casos, se supone que los datos siguen una distribución normal con una media constante y una varianza determinada. Estas técnicas intentan encontrar una relación matemática subyacente entre el sistema de entrada (señales de los sensores) y sus salidas (clases o descriptores).

Métodos no paramétricos: los métodos no paramétricos no tienen en cuenta la función de densidad probabilística para los datos de entrada y su aplicación es más genérica. Dentro de esta clasificación nos encontramos con las redes neuronales artificiales y los sistemas expertos.

Supervisado: en un método PARC de aprendizaje supervisado, un conjunto de datos de entrada conocido es sistemáticamente presentado al sistema, datos que son clasificados de acuerdo a los descriptores o clases determinadas previamente, y así, en una segunda etapa de identificación, datos de entrada no conocidos o nuevos se prueban con un conjunto de validación para mirar hasta que punto los métodos supervisados son capaces de predecir las correspondientes clases.

En la tabla 3.1 se muestran algunos de los principales algoritmos de reconocimiento de patrones que se pueden encontrar, entre ellos las redes neuronales fuzzy ARTMAP, PNN y los SVM empleados a lo largo de este proyecto.

A continuación se hace un estudio a fondo de las principales técnicas de reconocimiento de patrones que se han empleado en esta tesis como son las redes neuronales (fuzzy ARTMAP, PNN) y los Support Vector Machines (SVM).

Técnica	*Algoritmo*	Aprendizaje	Paramétrica	Aplicación
PCA	Lineal	No supervisado	No	Clasificación
PLS	Lineal	Supervisado	Si	Cuantificación
Feedforward-backpropagation ANN	Neuronal	Supervisado	No	Clasificación/ cuantificación
Fuzzy Art	Neuronal	No supervisado	No	Clasificación
Fuzzy Artmap	neuronal	Supervisado	No	clasificación
PNN	Neuronal	Supervisado	No	clasificación
SVM	No lineal	Supervisado	No	Clasificación, regresión

Tabla.3.1 Principales características de los algoritmos aplicados.

3.3 Redes neuronales

3.3.1 Definición

Existen numerosas formas de definir lo que son las redes neuronales, desde las definiciones cortas y genéricas hasta las que intentan explicar más detalladamente lo que significa “red neuronal” o “computación neuronal” [4].

ISBN:978-84-693-4070-7/DL:T-1167-2010

Las redes neuronales artificiales están compuestas por la interconexión masivamente en paralelo de elementos simples (usualmente adaptativos) y con organización jerárquica, cuya misión es interactuar con los objetos del mundo real de la misma manera que lo hace el sistema nervioso biológico.

Es necesario destacar que tales ordenadores neuronales no ejecutan las típicas instrucciones de máquina de los ordenadores digitales, a no ser que las utilicen para emular el comportamiento de las redes neuronales físicas. En principio la operación de procesos básicos realizada por todos los procesadores elementales es una operación análoga de transformación de sus señales de entrada.

En las redes neuronales biológicas, las células neuronales (neuronas) corresponden a los elementos de procesos. Las interconexiones se realizan por medio de las ramas de salida (axones) que producen un número de conexiones (sinapsis) con otras neuronas. Las redes neuronales son sistemas de simples elementos de proceso muy interconectados.

Una peculiaridad de las redes neuronales biológicas es el número elevado de procesadores o neuronas: en todo el sistema nervioso central hay del orden de 10^{11} neuronas, pero el número de interconexiones es aun mucho más grande, probablemente sobre 10^{15}.

3.3.2 Ventajas de las redes neuronales

Debido a su constitución y a su fundamento, las redes neuronales artificiales presentan un gran número de características similares a las del cerebro. Por ejemplo, son capaces de aprender de la experiencia, de generalizar de casos anteriores a nuevos casos, de abstraer características esenciales a partir de entradas que representan información irrelevante, etc. En general ofrecen numerosas ventajas, entre las que se incluyen:

- El aprendizaje adaptativo.
- La Auto-organización.

- La tolerancia a errores.
- La operación en tiempo real.
- La Fácil inserción a las nuevas tecnologías.

3.3.2.1 Aprendizaje adaptativo

La capacidad de aprendizaje adaptativo es una de las características más atractivas de las redes neuronales, ya que aprenden a realizar ciertas tareas mediante un entrenamiento con ejemplos ilustrativos. Como las redes neuronales pueden aprender a diferenciar patrones mediante ejemplos y entrenamiento, no es necesario que elaboremos modelos a priori ni necesitamos especificar funciones de distribución de probabilidad.

Las redes neuronales son sistemas dinámicos autoadaptativos. Son adaptables debido a la capacidad de autoajustar los elementos procesales (neuronas) que componen el sistema. Además de adaptativos son dinámicos ya que son capaces de estar constantemente cambiando para adaptarse a las nuevas condiciones, incluso en la fase de operación.

En el proceso de aprendizaje, los pesos (conexiones ponderadas) de las neuronas se ajustan de forma que se aprenda la relación entre unas señales de entrada y los resultados deseados. Una red neuronal no necesita un algoritmo específico para resolver cada problema, ya que puede generar su propia distribución de los pesos (enlaces) mediante el aprendizaje. También existen redes que continúan aprendiendo a lo largo de su vida, después de completar el periodo inicial de entrenamiento.

La función del diseñador es únicamente escoger la arquitectura neuronal apropiada. El usuario no es necesario que sepa como la red aprende a discriminar aunque siempre es conveniente que escoja el algoritmo de aprendizaje adecuado junto a un conjunto de datos estadísticamente representativo.

ISBN:978-84-693-4070-7/DL:T-1167-2010

3.3.2.2 Auto-organización

Las redes neuronales utilizan su capacidad de aprendizaje adaptativo para auto organizar la información que reciben durante el aprendizaje ó durante la fase de operación. Mientras que el aprendizaje es la modificación de los pesos de conexión de cada elemento procesal, la auto-organización consiste en la modificación de la red neuronal completa.

Esta característica es muy importante cuando se tiene que solucionar problemas en los cuales la información de entrada es poco clara; también, permite que el sistema entregue una solución incluso cuando la información de entrada esta especificada de forma incompleta.

3.3.2.3 Tolerancia a los errores

Las redes neuronales son los primeros métodos computacionales con la capacidad inherente de tolerancia a errores. Comparados con los sistemas computacionales tradicionales, los cuales pierden su funcionalidad cuando presentan un pequeño error de memoria, en las redes neuronales, si se produce un error en un pequeño número de neuronas, aunque el comportamiento del sistema se vea influenciado, el sistema no presenta una caída repentina.

Hay dos aspectos diferentes con respecto a la tolerancia de errores: primero, la red pueden aprender a reconocer patrones con información distorsionada o incompleta (tolerancia a los errores respecto a los datos) o bien pueden seguir realizando su función (con cierta degradación) aunque se destruya una parte de la red (tolerancia a fallos en el funcionamiento interno de la red).

La razón principal por la que las redes neuronales son tolerantes a los errores es el hecho de tener la información distribuida en las conexiones entre neuronas, lo que conlleva un cierto grado de redundancia en este tipo de almacenamiento. La mayoría de los ordenadores algorítmicos y sistemas de recuperación de datos almacenan cada pieza de información en un espacio único, localizado y direccionado. Las redes neuronales almacenan información no localizada. Por tanto, la mayoría de las

interconexiones entre los nodos de la red tendrán unos valores en función de los estímulos recibidos, lo que genera un patrón de salida que representa la información almacenada.

3.3.2.4 Operación en tiempo real

En muchas aplicaciones es necesario procesar una cantidad ingente de datos en un espacio corto de tiempo. Las redes neuronales se adaptan bien a una implementación paralela. Las redes neuronales suelen requerir de una carga de computación asimétrica entre el proceso de entrenamiento y el de evaluación. Para que la mayoría de las redes la computación en la fase de operación permite el trabajar en tiempo real, ya que la necesidad de cambio en los pesos de las conexiones es mínimo. Por este motivo las redes neuronales suelen ser una de las mejores alternativas en el reconocimiento y clasificación de patrones en tiempo real.

3.3.2.5 Fácil inserción a las nuevas tecnologías

Una red neuronal puede ser rápidamente entrenada, comprobada, verificada y trasladada a una implementación hardware de bajo coste, lo cual prueba la facilidad de insertar redes neuronales para aplicaciones especificas dadas de sistemas existentes. De esta manera, las redes neuronales se pueden utilizar para mejorar sistemas de forma incremental, y cada paso puede ser evaluado antes de pasar a un desarrollo más amplio.

3.3.3 Aplicaciones de las redes neuronales.

Las redes neuronales son una tecnología computacional emergente que puede utilizarse en un gran número y variedad de aplicaciones, tanto comerciales como militares. Se pueden desarrollar redes neuronales en un periodo de tiempo razonable y pueden desarrollar tareas concretas mejor que otras tecnologías más convencionales, incluyendo a los sistemas expertos. Cuando se implementan

ISBN:978-84-693-4070-7/DL:T-1167-2010

mediante hardware (redes neuronales en chips VLSI) presentan una alta tolerancia a errores del sistema y proporcionan un alto grado de paralelismo en el proceso de datos. Además, se hará posible insertar redes neuronales de bajo costo en sistemas existentes y recientemente desarrollados.

En la actualidad una gran cantidad de los sistemas de olfato electrónico comerciales incorporan sistemas de reconocimiento de patrones basados en técnicas de la estadística multivariante, análisis de componentes principales, análisis mediante funciones discriminantes y en redes neuronales. En el contexto de los sistemas olfativos artificiales, las redes neuronales se encargan de realizar las tareas de clasificar y/o cuantificar muestras.

En los siguientes apartados se describen con más detalle dos arquitecturas de redes neuronales que han sido ampliamente utilizadas en el contexto de los sistemas de olfativo electrónicos. Estas redes han sido implementadas y juegan un papel importante en el desarrollo de esta tesis.

3.3.4 Redes FUZZY ART

3.3.4.1 Introducción

La teoría de la resonancia adaptativa (ART) fue introducida como una teoría que intentaba emular la manera en como el cerebro humano procesa la información. Desde entonces, esta teoría 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 arbitrarias con un ritmo de aprendizaje rápido o lento. Dentro de estos algoritmos se pueden destacar el ART, ART2 y ART3. [5,6,7].

Fuzzy ART es una evolución del algoritmo ART1. Este último es capaz de categorizar de forma estable entradas arbritarias 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 (∩) y unión (U) de ART1 por los operadores MIN (Λ) y MAX (V),

respectivamente, de la teoría de lógica difusa. Este cambio, con la ayuda de la codificación complementaria, 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 3.1 se muestra un esquema del algoritmo. Su modo de funcionamiento es simple: cada vez que la red recibe un nuevo vector de entrada V reacciona activando uno y sólo uno de los nodos de salida. Cada uno de estos nodos representa una de las diferentes clases 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.

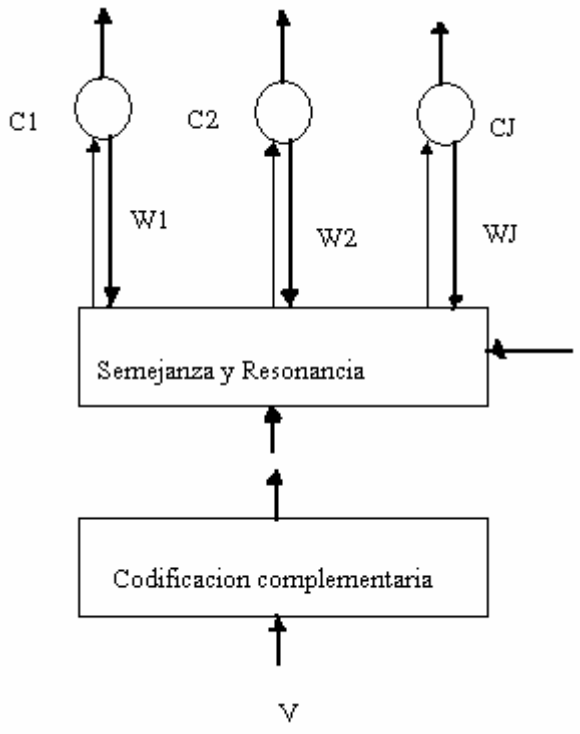

Figura 3.1 esquema de la red Fuzzy Art

ISBN:978-84-693-4070-7/DL:T-1167-2010

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 β, siendo su valor igual a la unidad para un aprendizaje rápido e igual a cero en caso de que no se deba aprender más.

3.3.4.2 Algoritmo

Incluimos, a continuación, una descripción esquemática del algoritmo:

- Vector de entrada: cada uno de los vectores de entrada V es un vector M-dimensionalidad 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 $I_{J+M}= I-I_J$.
- Vector de pesos del nodo de salida j (categoría j): W_J. Inicialmente, $W_{J1}= W_{J2}=W_{J2M}=1$.
- Velocidad de aprendizaje, β entre [0 1]. Aprendizaje rápido, $\beta =1$; aprendizaje lento, $\beta<<1$; sin aprendizaje, $\beta=0$.
- Parámetro de vigilancia: ρ cercano a cero implica menos categorías al agrupar con criterios de semejanza poco exigentes, ρ cercano a uno implica muchas clases, cada una con pocos miembros pero muy parecidos entre si.

- 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 $T_J(V)$ como indica la ecuación.

$$T_J = \left| I \Lambda W_J \right| / \alpha + \left| W_J \right| \tag{3.1}$$

Donde el operador AND (Λ) en lógica difusa se define como:

$$A \Lambda B = \min (A,B) \tag{3.2}$$

y la norma $\left| \right|$ se define como:

$$\left| I \right| = \sum^{2M} I_i \tag{3.3}$$

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

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

$$\frac{\left| \boldsymbol{I} \wedge \boldsymbol{W}_J \right|}{\left| \boldsymbol{I} \right|} < \rho \tag{3.4}$$

En ese caso, se activa la 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 desigualdad se produce un reset: el sistema desactiva temporalmente la categoría j y vuelve a escoger una categoría siguiendo el criterio de máxima semejanza. Si ninguna categoría "resuena", se crea un nuevo nodo para el vector de entrada V.

ISBN:978-84-693-4070-7/DL:T-1167-2010

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

$$\boldsymbol{W}_J^{(new)} = \beta(\boldsymbol{I} \wedge \boldsymbol{W}_J^{(old)}) + (1-\beta)\boldsymbol{W}_J^{(old)}; \qquad 0 < \beta \leq 1 \tag{3.5}$$

Si se quiere un aprendizaje rápido, se utiliza una β= 1, 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.

3.3.5 Redes 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 [8].

La red fuzzy ARTMAP proviene de la red ARTMAP 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 (componentes comprendidos entre O y 1) de la red binaria ARTMAP.

Básicamente, una red fuzzy ARTMAP esta formada por dos redes Fuzzy ART conectadas entre si por un vector de relaciones denominado “mapfield o memoria asociativa”. A una de las dos redes (la que denominamos 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 3.2 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 medidas que deban ser clasificadas conjuntamente enviaran a la red B codificaciones idénticas. Además, cualquier vector de codificación diferente, por parecido que sea el resto, debe ser detectado y debe activar una neurona de salida diferente en la red B.

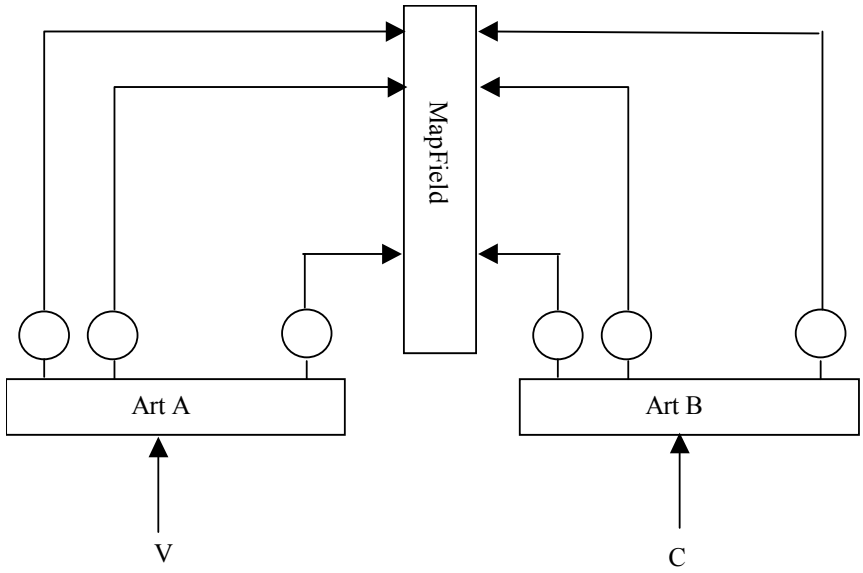

Figura 3.2 esquema general de una red Fuzzy ARTMAP

Cada vez que se suministra una medida de entrenamiento, la red A activa un nodo y la 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 se le asocia un nodo en B. cabe destacar que los nodos B normalmente serán 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 se creara una nueva y el mapa le asignara como imagen el nodo B activado.

En definitiva, el valor de vigilancia A solo se incrementa lo estrictamente necesario para que la red A separe en nodos diferentes las medidas que deben estar

ISBN:978-84-693-4070-7/DL:T-1167-2010

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.

La red Fuzzy ARTMAP presenta múltiples ventajas que la hacen muy interesante para las aplicaciones con narices electrónicas. 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 costosa 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 el proceso de aprendizaje.
- Aprendizaje continúo 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 a luz sobre los procesos internos y como influyen en la categorización de resultados.

Sin embargo, su implementación práctica presenta un problema que debe ser tratado con sumo cuidado. Tal y como se ha comentado con anterioridad, la red aprende el 100 % de las medidas del conjunto de entrenamiento sacrificando lo

mínimo posible la generalización. Sin embargo, la presencia de outliers o medidas erróneas en el conjunto de entrenamiento puede requerir un incremento del valor de vigilancia excesivo, lo cual perjudicara seriamente la capacidad de generalizar de la red. Este problema es una de las razones por las que este tipo de algoritmo se ha utilizado poco en EN ya que en este tipo de aplicaciones es sumamente difícil identificar medidas erróneas dada la baja repetitividad de las señales de los sensores. Cabe aclarar que se ha utilizado una modificación de dicha red, propuesta e implementada por Jesús Brezmes para evitar este problema [8].

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 le suministra el vector de entrada y la red clasifica dicha medida siguiendo los criterios que ha aprendido en la fase de entrenamiento.

3.3.6 Red PNN (Probabilistic neural networks)

Esta red es muy utilizada en problemas de clasificación. La red consta de dos capas, una red de neuronas de base radial con un número de neuronas igual al número de vectores de entrenamiento y una capa competitiva de neuronas cuyo número es igual al número de categorías consideradas en el problema de clasificación. En la figura 3.3 se muestra un esquema general de esta arquitectura de red.

Cada neurona de base radial (de la capa de entrada) almacena como pesos uno de los vectores de entrenamiento. En la fase de clasificación o reconocimiento, se presenta un nuevo vector de entrada. En la primera capa se calculan las distancias Euclídeas entre el vector de entrada y los pesos de cada neurona. En general, en la entrada de cada neurona de base radial se obtiene un escalar indicativo del parecido entre el vector a clasificar y los pesos de dicha neurona. Si para una neurona el escalar es cero, entonces el vector a clasificar ha resultado ser idéntico al vector de

ISBN:978-84-693-4070-7/DL:T-1167-2010

pesos de dicha neurona. Los escalares resultantes del cálculo de la distancia son multiplicados por un escalar denominado 'spread' que es idéntico para todas las neuronas de base radial de la red e introducidos como argumento en una función de tipo Gaussiana denominada 'radbas'.

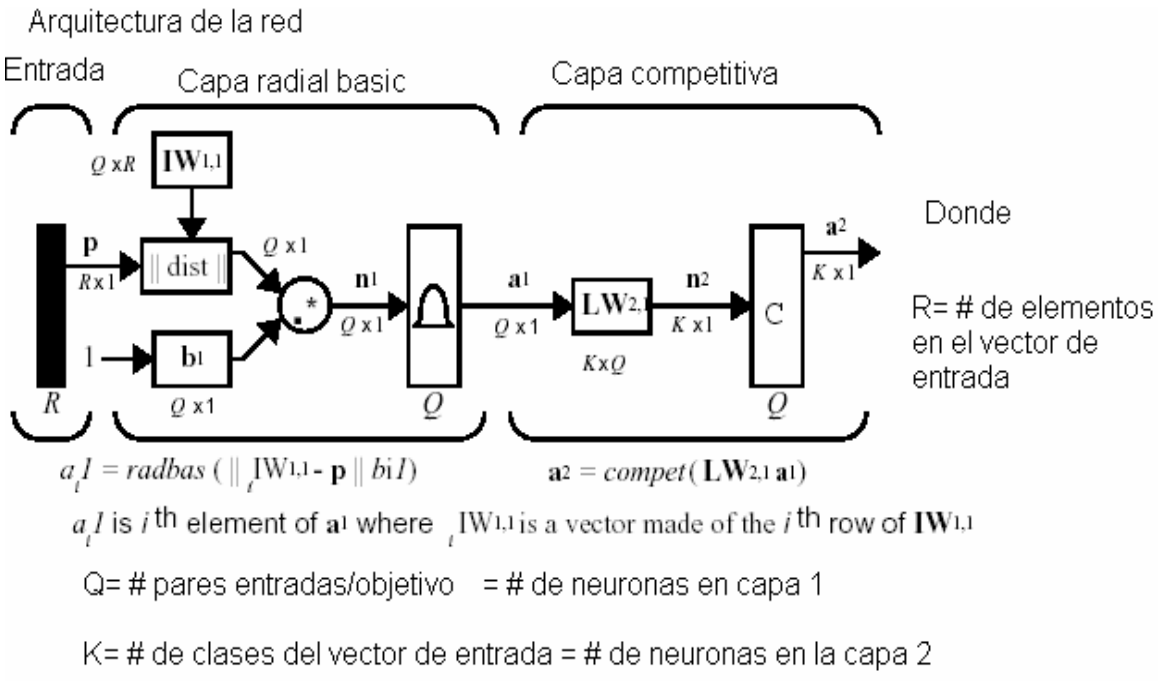

Figura 3.3 Esquema de la red PNN (extraído de [9])

La expresión de dicha función (donde se ha denominado n al argumento) es:

$$radbas(n) = e^{-n^2} \tag{3.6}$$

La figura 3.4 muestra el aspecto de dicha función que adquiere su valor máximo (igual a la unidad) cuando su argumento vale 0. Por lo tanto, dado un vector nuevo que deba ser clasificado, la neurona de base radial que produzca un máximo a su

salida será aquella cuyos pesos más se parezcan al vector de entrada. La función del parámetro 'spread' es determinar la anchura de la campana de Gauss.

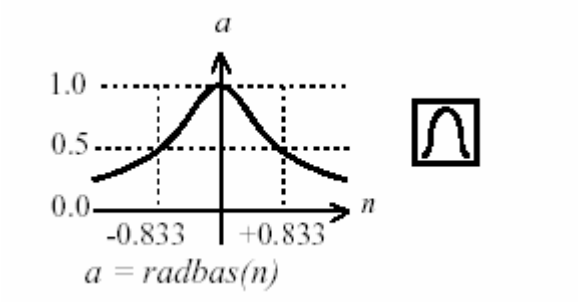

Figura 3.4. *Función de transferencia de una neurona de base radial*

La segunda capa suma las contribuciones de las diferentes neuronas para generar un vector de dimensión igual al número de categorías. Este proceso es muy sencillo. Imaginemos que cuando resulta ganadora la neurona de base radial *i* esto implique que el vector de entrada pertenece a la categoría *j*. Entonces el peso del enlace entre la neurona de base radial *i* y la neurona competitiva *j* es igual a la unidad mientras que los pesos que enlazan a la neurona *i* con las demás neuronas competitivas (las otras categorías) son cero. El elemento de la posición i-ésima del vector resultante de sumar las contribuciones de las neuronas de base radial puede ser interpretado como la probabilidad de que el vector de entrada pertenezca a la categoría i-ésima. Finalmente, una función de transferencia "competitiva" sobre la salida de la segunda capa escoge el máximo entre todas las probabilidades, y produce un uno para esa clase y un cero para todas las demás clases [9].

La red probabilística (PNN) realiza un entrenamiento sumamente rápido puesto que el proceso de ajuste de los pesos no es iterativo. Se ha demostrado que esta red se comporta como el algoritmo k-NN (k-nearest neighbour), si se escoge un valor para el 'spread' suficientemente bajo (típicamente de 0.1) [9].

ISBN:978-84-693-4070-7/DL:T-1167-2010

3.4 Support Vector Machines

3.4.1 Introducción

Los SUPPORT VECTOR MACHINES (SVM) son una familia de algoritmos de clasificación y regresión desarrollados por Vapnik y colaboradores [10,11,12] y reconocidos por la comunidad científica a principio de los años noventa. Hoy en día, los SVM muestran resultados comparables o mejores que los obtenidos con redes neuronales y otros modelos estadísticos en una gran cantidad de problemas como son entre otros, visión por computador, reconocimientos de patrones y minería de datos (data mining) [13].

La comunidad de los sistemas de olfato artificial comenzó a explorar el uso de SVM muy recientemente (el primer artículo publicado apareció en el año 2003) [14,15], y el número de contribuciones presentados hasta el día de hoy sigue siendo relativamente bajo [16-23]. Los SVM poseen algunas características que los hacen muy atractivos en el contexto de los sistemas olfativos y multisensoriales. Entre estas características destacan las siguientes:

- Los SVM son no paramétricos: Los parámetros del modelo no son predefinidos y su número depende del conjunto de datos de entrenamiento disponible. Lo que se buscaría idealmente durante el proceso de entrenamiento es emparejar la capacidad del modelo a la complejidad de los datos. Esta propiedad es compartida por los SVM con las redes neuronales como por ejemplo las basadas en el perceptron multicapa (MLP) o la red basada en funciones en base radial (RBF).
- A diferencia de las redes neuronales clásicas, donde la estructura del modelo ya viene preseleccionada (es decir, el número de neuronas ocultas está predeterminado de antemano), el intervalo de confianza (error de estimación) está fijado y el error de entrenamiento se minimiza, con los SVM el error de entrenamiento se fija mientras que el intervalo de confidencia se minimiza. Por lo tanto, a diferencia de los algoritmos de adaptación clásicos (MLP y RBF), los SVM realizan la minimización estructural del riesgo en el marco

de la teoría de regulación. Esto significa que los SVM reducen al mínimo la dimensión del modelo y funcionan bien con datos desconocidos. En otras palabras, los SVM muestran una óptima habilidad para la generalización y son menos propensas a sufrir sobreentreno (over-fitting) que otros modelos más conocidos.

- La tarea de un SVM consiste en encontrar un hiperplano óptimo lineal que separe las clases. Esto es equivalente a resolver un problema clásico de programación cuadrática. Mientras que para las clases separables linealmente el hiperplano se calcula en el espacio de entrada original, para clases de separación no lineales los SVM aplican una transformación no lineal del espacio de entrada en un espacio multidimensional donde las clases sean linealmente separables y poder calcular, de esta forma, el hiperplano de separación. Esta transformación puede realizarse mediante diferentes mapeados tales como funciones polimoniales, RBF (gausiana o multicuadratica) o funciones spline. El mapeo no lineal o kernel permiten realizar todos los cálculos en el espacio de entrada. Por otro lado, los SVM tienen unos requisitos computacionales bajos y la alta dimensionalidad del espacio de entrada no es un problema. Esto es ventajoso porque el sistema puede ser reentrenado periódicamente.
- En un principio, los SVM se han desarrollado para resolver problemas de clasificación, aunque también pueden ser aplicados con éxito a casos de regresión [24], por lo que, esta técnica ha sido usada en esta tesis para construir modelos que permiten estimar las concentraciones de cada contaminante en una muestra gaseosa.

Determinar la dimensión correcta del espacio de entrada es un aspecto importante que debería ser estudiado cuidadosamente al construir modelos de regresión y clasificación en el contexto de las narices electrónicas y los analizadores de gases. Esto es particularmente cierto cuando se considera el uso de SVM como modelos de reconocimiento de patrones (PARC).

ISBN:978-84-693-4070-7/DL:T-1167-2010

A continuación se estudia mas a fondo los conceptos relacionados a los modelos SVM tanto para clasificación como para regresión, [25,26] que serán empleados más adelante con nuestros conjuntos de datos para realizar procesos de selección de variables.

3.4.2 SVM para clasificación:

El principio de los SVM consiste en construir un hiperplano óptimo lineal que permita separar las clases procurando que:

- Los vectores que pertenecen a las distintas clases se encuentran en distintos lados del hiperplano.
- La mínima distancia entre los vectores y el hiperplano (el margen) sea la máxima posible.

Si los datos no son linealmente separables se busca proyectar los datos del espacio de entrada (que pertenece a dos clases diferentes) en un espacio de mayor dimensión llamado espacio de características de modo que los datos se vuelvan linealmente separables.

3.4.2.1 Caso linealmente separable

Se dice que un clasificador es lineal cuando es posible expresar su función de decisión por una función lineal en x. En consecuencia, suponemos que los ejemplos se proporcionan en formato vectorial. Nuestro espacio de entrada X corresponde a $\Re^n$ dónde n es el número de componentes de los vectores que contienen los datos. Sin pérdida de generalidad, la función de clasificación puede expresarse de la siguiente manera:

$$f(x) = \langle \omega, x \rangle + b \tag{3.7}$$

$$= \sum_{i=1}^{n} \omega_i x_i + b \tag{3.8}$$

Donde $\omega \in \Re^n$ y $b \in \Re$ son los parámetros que definen estos hiperplanos, y $x \in \Re^n$ es una variable.

Para decidir a que categoría pertenece una medida, basta con tomar la señal de la función de decisión: $y = \text{sgn}(f(\tilde{x}))$

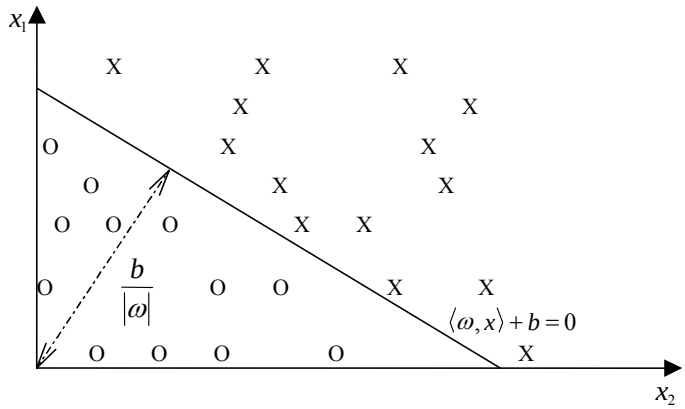

Figura. 3.5: Representación en el hiperplano que corresponde a la función de decisión de un clasificador lineal.

Geométricamente, eso equivale a considerar un hiperplano que es el lugar de los puntos x en los que se cumple:

$$\langle w, x \rangle + b = 0 \tag{3.9}$$

Al orientar el hiperplano, la regla de decisión corresponde a observar en que lado del hiperplano se encuentra la medida $\tilde{x}$. La figura 3.5 representa la situación en $\Re^n$

Se ve que el vector ω define la pendiente del hiperplano ya que es perpendicular al hiperplano. El término b, por su parte, permite determinar cual es el hiperplano de los infinitos paralelos que existen.

ISBN:978-84-693-4070-7/DL:T-1167-2010

3.4.2.2 Margen del hiperplano y solución del problema

El margen es la distancia mínima entre las muestras del conjunto de aprendizaje y la frontera de decisión. El margen puede medirse gracias al vector de pesos ω. Puesto que suponemos que las muestras son separables, se puede redefinir ω y b para lo que las muestras x (support vectors) más próximas al hiperplano satisfagan la ecuación $\left|\langle w, x\rangle + b\right| = 1$.

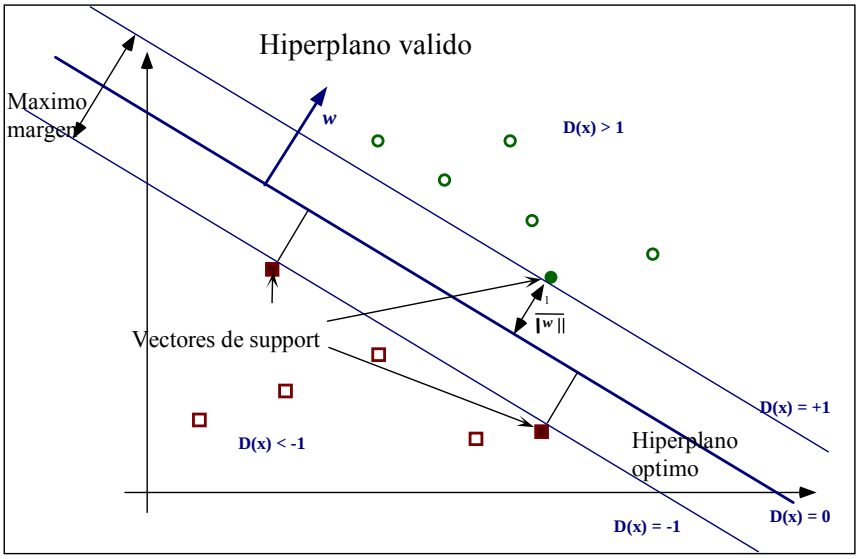

Figura. 3.6: Representación de $\Re^2$ en el hiperplano que corresponde a la función de decisión de un clasificador lineal con el mayor margen.

Dado un conjunto de puntos de aprendizaje $(x_i, y_i), i \in [1, l]$ donde cada $x_i \in \Re^n$ y y_i a $\{-1,1\}$, y_i define la clase de un ejemplo dado (de entre dos posibles). El objetivo de los SVM es encontrar un hiperplano que permita separar el conjunto de aprendizaje de modo que todos los puntos de una misma clase estén a un mismo lado

del hiperplano. Eso equivale a encontrar un hiperplano $f(x)$ definido por $\omega \in \Re^n$ b$\in\Re$ tal que:

$$y_i\left(\langle \omega, x_i \rangle + b\right) \geq 1 \text{ con } i \in [1,l] \tag{3.10}$$

Entre el conjunto de los posibles hiperplanos que satisfacen estas condiciones, los SVM buscan aquel que maximice la distancia entre el hiperplano y los puntos más próximos a cada clase (Figura 3.6), dado que se definió esta distancia como $2/\|\omega\|$ [27]. Por lo que el problema global equivale a encontrar:

$$min_{\omega,b} \frac{1}{2}\|\omega\|^2 \tag{3.11}$$

$$\text{Tal que } y_i f(x_i) \geq 1 \quad i \in [1,l] \tag{3.12}$$

Este problema puede solucionarse utilizando los multiplicadores de Lagrange α:

$$\omega = \sum_{i=1}^{l} \alpha_i^* y_i x_i \tag{3.13}$$

donde los α_i^* son la solución del siguiente problema

$$min_{\alpha} \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle - \sum_{i=1}^{l} \alpha_i \tag{3.14}$$

$$\text{Tal que } \sum_{i=1}^{l} \alpha_i y_i = 0 \tag{3.15}$$

$$\alpha_i \geq 0 \quad \forall i$$

Este último es un problema de optimización cuadrática standard. Finalmente el hiperplano de decisión estaría dado por:

ISBN:978-84-693-4070-7/DL:T-1167-2010

$$f(x) = \sum_{i}^{l} \alpha_i y_i \langle x_i, x \rangle + b \qquad (3.16)$$

Dónde b es el multiplicador de Lagrange asociado a la restricción $\sum_{i=1}^{l} \alpha_i y_i = 0$ del problema dual.

3.4.2.3 Caso no lineal

En la mayoría de los problemas reales, los datos no son linealmente separables, por lo que es necesario abordar este problema con una nueva estrategia:

Si se proyectan los puntos de aprendizaje en un nuevo espacio H gracias a una función $\varphi(x)$ y en el se aplica el mismo método de máxima optimización en el espacio *H*, casi siempre el problema se vuelve linealmente separable.

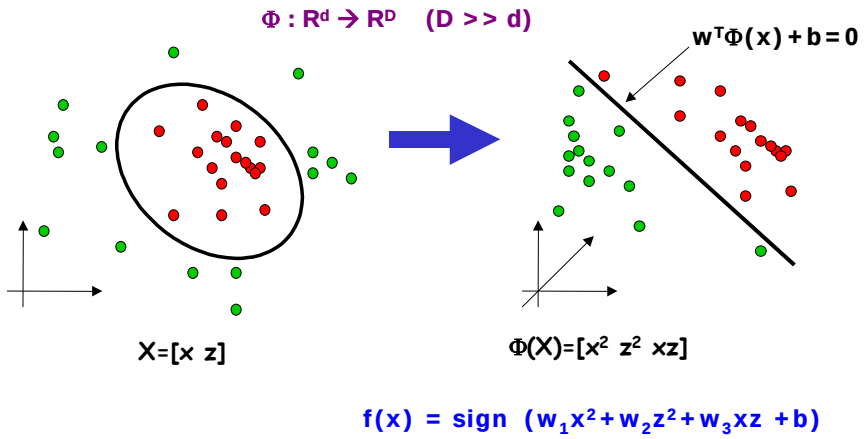

Figura 3.7: Ejemplo de proyección en un espacio de redescripción de gran dimensión donde el problema se vuelve linealmente separable.

En efecto, en cuanto más grande sea la dimensión del espacio de descripción, mayor es la probabilidad de poder encontrar un hiperplano separador entre los ejemplos y los contraejemplos. Al transformar el espacio de entrada en un espacio dimensionalmente superior, (incluso de dimensión infinita), resulta posible utilizar de nuevo la separación mediante hiperplanos (Figura 3.7).

Denotemos a Φ como una transformación no lineal del espacio de entrada X en un espacio de dimensión superior $\Phi(X)$:

$$\text{Es decir } \Phi \quad \Re^n \rightarrow H$$

Así el caso no-lineal se obtiene fácilmente sustituyendo $\langle x_i, x_j \rangle$ por $k(x_i, x_j) = \langle \Phi(x_i, x_j) \rangle$ dónde k es una función llamada "función núcleo" ("kernel"). Por lo tanto se pueden efectuar todos los cálculos utilizando k, sin tener que transformar los datos por la función φ, con lo que no es estrictamente necesario el conocer la función φ. En la clasificación de un nuevo dato x, se calcula la señal de la función f como:

$$f(x) = \sum_i^l \alpha_i y_i k\langle x_i, x \rangle + b \tag{3.17}$$

Las funciones núcleos aceptables deben cumplir la condición de Mercer. El cumplimiento de la condición de Mercer garantiza que el problema cuadrático tenga solución. En la práctica, los núcleos más utilizados son:

- Los núcleos polinomiales de orden p

$$k(x, y) = (\langle x, y \rangle + 1)^p \tag{3.18}$$

- Los núcleos de Gauss (Radial Basis Function (RBF)) de anchura de banda σ

ISBN:978-84-693-4070-7/DL:T-1167-2010

$$k(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}} \tag{3.19}$$

3.4.2.4: Caso no separable

Cuando el conjunto de aprendizaje no es separable, es necesario introducir variables de relajación en la definición de las restricciones. A cada dato se le asocia una nueva variable ξ_i que nos indicara si el dato está del lado correcto o no del separador.

Un dato x_i esta bien clasificado si $\xi_i = 0$. Si $\xi_i \neq 0$ es decir, si x_i esta mal clasificado, entonces $\xi_i \geq 1$; así pues, ξ_i indica hasta qué punto el dato x_i esta del lado equivocado. El problema se centra en la búsqueda del hiperplano que implica el margen más grande y el número de error más pequeño. La función objetiva pasaría a ser:

$$min_{\omega,b,\xi_i} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l}\xi_i^2 \tag{3.20}$$

$$\text{Tal que} \qquad y_i f(x_i) \geq 1 - \xi_i \qquad i \in [1,l] \tag{3.21}$$

Dónde ξ_i es la variable de relajación de una dificultad y C un coeficiente de penalización de la relajación. El tratamiento de este problema se hace de manera simple poniendo de manifiesto que la adición de la penalización cuadrática de los puntos mal clasificados equivale a tratar el caso separable sustituyendo $k(x_i, x_j)$ por:

$$k(x_i, x_j) + \frac{1}{C}\delta_{i,j} \tag{3.22}$$

donde $\delta_{i,j}$ es el simbolo de Kronecker.

3.4.3 SVM multiclase

Inicialmente los SVM fueron desarrollados para resolver problemas binarios. Sin embargo, existen diferentes estrategias que permiten desarrollar técnicas de SVM para resolver problemas de N clases [28]. Entre ellas podemos destacar las siguientes:

- *Uno contra todos:* en esta estrategia, se construyen N modelos SVM. El *i-th* SVM es entrenado con todas las muestras de entrenamiento de la clase *i- th* con etiquetas (labels, valores) positivos, y todas las otras muestras de entrenamiento con labels negativos. Una nueva medida *x* pertenece a la clase que tiene los valores altos de la función indicador.

$$\text{Clase de } x \equiv \underbrace{\arg\ \max}_{i=1,..,N}\left((\omega^i)^t \Phi(x) + b^i\right) \qquad (3.23)$$

- *Uno contra uno*: este método construye N(N -1) / 2 clasificadores donde cada uno es entrenado usando patrones de dos clases. Para clasificar una medida se implementa un sistema de voto. Si la función indicador $\text{sign}\left((\omega^{ij})^t \Phi(x) + b^{ij}\right)$ dice que *x* pertenece a la clase *i*, entonces el voto para la clase *i*, se incrementa en uno, si no, el voto para la clase *j* se incrementa en uno. En caso de igual número de votos entre diferentes clases, se selecciona aquella con índice más pequeño.
- *Gráfico acíclico directo (Direct acyclic graph): esta* fase es similar al método uno contra uno. En la fase de prueba se usa un gráfico acíclico binario (rooted binary directed acyclic graph) con N(N-1)/ 2 nodos internos y N hojas (leaves). Cada nodo es un SVM binario de clases *i-th* y *j-th*. Con una nueva medida x, empezando el nodo en la raíz, se evalúa la función indicador binaria. Entonces se mueve a la derecha o hacia la izquierda dependiendo del resultado. El método procede hasta que se alcanza el nodo de una hoja, que indica la clase predecida. El tiempo de prueba para este método es menor que el ejecutado en el uno contra uno.

ISBN:978-84-693-4070-7/DL:T-1167-2010

3.4.4 Regresión mediante SVM's:

Para el propósito de la regresión, se estudia la dependencia funcional de la variable de salida $y \in \Re$ sobre una variable de entrada *n*-dimensional. Si consideramos el caso de la regresión lineal, un hiperplano de regresión lineal $f(x, \omega)$, se define como:

$$f(x, \omega) = \omega^t x + b \tag{3.24}$$

En este caso se emplea el error de aproximación en lugar del margen usado en la clasificación. La función de pérdida lineal deVapnik's con insensibilidad ε se define como:

$$e(x, y, f) = \left|y - f(x, \omega)\right|_\varepsilon = \begin{cases} 0 & \text{if } \left|y - f(x, \omega)\right| \le \varepsilon \\ \left|y - f(x, \omega)\right| - \varepsilon & \text{si no} \end{cases} \tag{3.25}$$

Por lo tanto, la pérdida es igual a cero si la diferencia entre el valor actual y el predicho de *y* es menor que ε. Como en la clasificación, el método procura minimizar el riesgo empírico y $\frac{1}{2}\omega^t\omega$ simultáneamente, por lo que el hiperplano es construido minimizando:

$$\frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{p}\left|y_i - f(x_i, \omega)\right|_\varepsilon \equiv \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{p}\left(\xi_i + \xi_i^*\right) \tag{3.26}$$

teniendo en cuenta las siguientes condiciones

$$\begin{cases} y_i - \omega^t x_i - b \le \varepsilon + \xi_i \\ \omega^t x_i + b - y_i \le \varepsilon + \xi_i^* \\ \varepsilon, \xi_i, \xi_i^* \ge 0 \end{cases}$$

Donde ξ_i y ξ_i^* son, respectivamente, variables "slack" para medidas por encima y por debajo de la zona de insensibilidad ε. La figura 3.8 ilustra el uso de la función de pérdida insensible ε en support vectors de regresión aplicando un procedimiento similar al usado para la clasificación. El problema de optimización puede ser resuelto reduciendo al mínimo el Lagrangiano dual:

$$L_d(\alpha_i, \alpha_i^*) = \frac{1}{2}\sum_{i,j=1}^{p}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)x_i^t x_j + \sum_{i=1}^{p}(\varepsilon - y_i)\alpha_i + \sum_{i=1}^{p}(\varepsilon + y_i)\alpha_i^* \quad (3.27)$$

Teniendo en cuenta que:

$$\sum_{i=1}^{p}(\alpha_i - \alpha_i^*) = 0\,,\ 0 \le \alpha_i \le \mathrm{C},\ 0 \le \alpha_i^* \le \mathrm{C}. \quad (3.28)$$

Mediante el cálculo de los multiplicadores de Lagrange α_i y α_i^* se encuentra el vector de peso óptimo del hiperplano de regresión, ω_s, ,:

$$\omega_s = \sum_{i=1}^{p}(\alpha_i - \alpha_i^*)x_i \quad (3.29)$$

El valor bias b puede calcularse usando los vectores de entrenamiento que satisfagan

$y - f(x,\omega) = \pm\,\varepsilon$. A estos se les denomina support vectors libres.

$$\begin{aligned} b &= y_i - \omega^t x_i - \varepsilon \quad \text{for } 0 \le \alpha_i \le C \\ b &= y_i - \omega^t x_i + \varepsilon \quad \text{for } 0 \le \alpha_i^* \le C \end{aligned} \quad (3.30)$$

Por lo que el hiperplano óptimo de regresión es:

$$f(x, \omega) = \sum_{i=1}^{p}(\alpha_i - \alpha_i^*)x_i^t x + b \quad (3.31)$$

Cuando se considera una regresión no lineal, se debe proceder de forma similar a la vista en el caso de la clasificación no lineal, definiendo un mapping $\Phi(x)$ sobre un espacio dimensional alto. En un espacio Φ el algoritmo de aprendizaje podrá realizar una regresión lineal. Usando la función kernel $K(x_i,x_j)= \Phi^t(x_i)\,\Phi(x_j)$ tendremos:

ISBN:978-84-693-4070-7/DL:T-1167-2010

$$f(x,\omega) = \omega_s^t \Phi(x) + b = \sum_{i=1}^{p} (\alpha_i - \alpha_i^*) \Phi^t(x_i)\Phi(x) + b = \\ = \sum_{i=1}^{p} (\alpha_i - \alpha_i^*) K(x_i, x) + b \tag{3.32}$$

El termino bias *b* se calcula de la siguiente manera para cualquier support vector superior e inferior.

$$b = y_i - \sum_{j=1}^{FU} (\alpha_j - \alpha_j^*) K(x_i, x_j) - \varepsilon \qquad \text{para } 0 \le \alpha_i \le C \\ b = y_i - \sum_{j=1}^{FL} (\alpha_j - \alpha_j^*) K(x_i, x_j) + \varepsilon \qquad \text{para } 0 \le \alpha_i^* \le C \tag{3.33}$$

Donde FU y FL son él número de support vectors libres superiores e inferiores respectivamente.

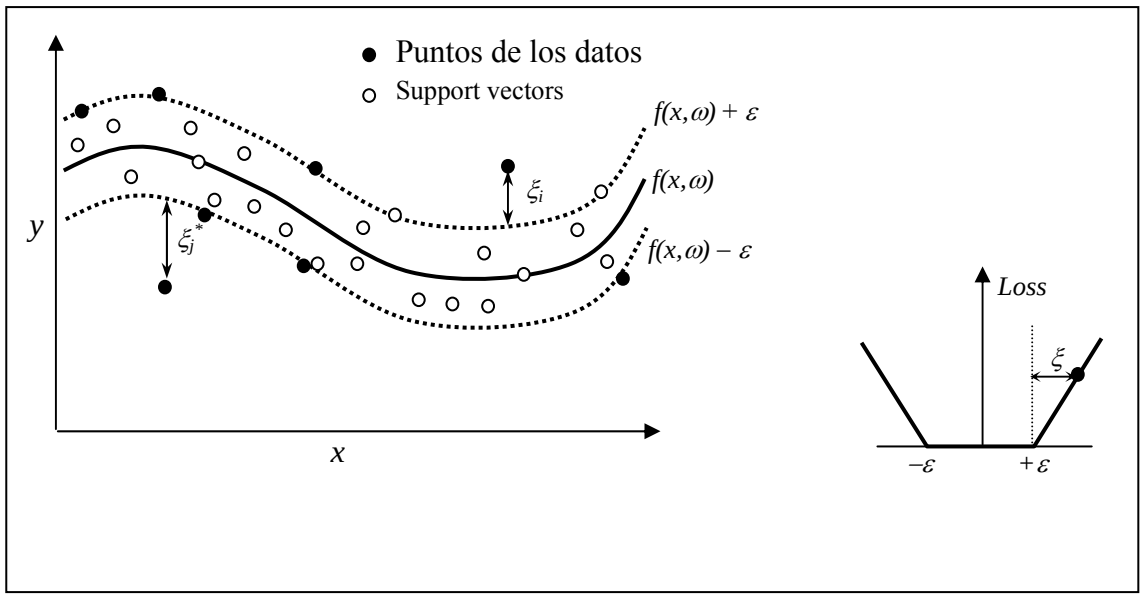

Figura.3.8: Parámetros usados en una regresión 1-dimensional y con una función de perdida ε- insensible. Los support vectors pueden aparecer solo en las fronteras o fuera de la zona ε-insensible.

3.5 Selección de variables

3.5.1 Introducción

De todo sistema de sensado se obtienen una serie de variables que pueden contribuir o no al correcto funcionamiento del sistema. Como norma general, y muy particularmente en el caso de los sistemas de olfato electrónico, no existe ninguna garantía de que aumentando el número de variables extraídas de la respuesta de los sensores se obtengan resultados más exactos. Algunas variables dan información útil, y otras proporcionan ruido no deseado. Por esta razón es necesario escoger las variables que serán utilizadas por los diferentes algoritmos de reconocimiento de patrones del modelo. Usando un criterio de selección de variables, la dimensionalidad de los datos puede reducirse sin perder información útil, y al mismo tiempo la información compuesta por ruido puede minimizarse. En definitiva, para estar seguro de que los resultados obtenidos sean buenos, es necesario seleccionar cuidadosamente las variables (parámetros) que se utilizaran junto a los algoritmos de reconocimiento de patrones que se deseen aplicar.

En el caso de los sistemas de olfato electrónico, la mayoría de las técnicas usadas para identificar cuáles son los parámetros que ayudan a discriminar entre gases simples o aromas complejos se basan en técnicas lineales como el análisis de los componentes principales (PCA) y los mínimos cuadrados parciales (PLS) [1]. Teniendo en cuenta que PLS y PCA se comportan de un modo similar desde el punto de vista de la selección de variables, limitamos aquí la explicación al método PCA.

PCA es una técnica de representación de la señal que genera proyecciones a lo largo de las direcciones de máxima varianza, que se definen por los primeros autovectores de la matriz de covarianza de la respuesta de los sensores. En este método, las variables iniciales x, se proyectan sobre los PC's obteniéndose las coordenadas de las medidas en el nuevo sistema de representación. A estas nuevas coordenadas se las denomina scores. Las nuevas variables (scores) pueden usarse entonces como entradas en el modelo de clasificador (por ejemplo una red neuronal).

ISBN:978-84-693-4070-7/DL:T-1167-2010

Como que el sistema de representación basado en los PCs es ortogonal, se resuelven problemas de colinealidad en la matriz de respuesta. Además como que los primeros PCs suelen capturar la mayor parte de la varianza útil en los datos, pocos PCs son suficientes para representar la información original. Esto conlleva una drástica reducción del número de variables de entrada que deben manejar los clasificadores. Sin embargo, los PCs son nuevas variables resultantes de la combinación lineal entre las variables originales (por ejemplo parámetros de la respuesta de los sensores). Por lo tanto no tienen sentido físico (o químico) directo y no presentan el interés que tiene eliminar directamente alguna variable original (por ruidosa, redundante, etc.). Es por eso que en esta tesis se han diseñado métodos de selección que trabajan directamente sobre variables originales y no sobre variables secundarias.

La selección de variables (SV) engloba a todo un conjunto de técnicas de reducción de la dimensionalidad de los datos a procesar [2]. La meta de la selección de variables es encontrar un subconjunto "óptimo" de variables que minimice la perdida de información y maximice la reducción de ruido. La estrategia para la SV más común consiste en evaluar cada variable individualmente y seleccionar aquellas variables que aportan información de mayor calidad. Desgraciadamente, este acercamiento ignora la redundancia o la sinergia entre variables y raramente encontrará un subconjunto óptimo. Ante esta situación uno puede tener la tentación de evaluar todos los posibles subconjuntos de variables y seleccionar el óptimo.

Cualquier procedimiento para la selección de las variables basa su funcionamiento en dos aspectos fundamentales: un criterio de selección y un procedimiento de búsqueda. Con suerte, la selección se haría investigando todos los posibles subconjuntos de las variables que se utilizan. Sin embargo, esto es impráctico, ya que por lo general implica investigar una cantidad casi infinita de combinaciones que requeriría un tiempo de cálculo inasumible en la mayoría de aplicaciones. Además, si el modelo escogido fuera no-lineal, por ejemplo una red neuronal artificial (ANN), los requisitos computacionales serian inabordables. Por lo tanto, el objetivo es así encontrar un criterio de selección simplificado y un procedimiento de búsqueda que proporcione resultados cercanos al óptimo global. Una vez aplicado el método, el

conjunto de variables que resulte seleccionado será usado luego como entradas al sistema de reconocimiento de patrones.

Para evitar la explosión exponencial de una búsqueda exhaustiva, se han desarrollado diferentes métodos que exploran el espacio de las variables de una manera más eficaz. Estas estrategias de búsqueda pueden agruparse en tres grandes categorías: exponenciales, secuenciales (deterministas), y aleatorias (estocásticas).

Las técnicas exponenciales realizan una búsqueda cuya complejidad crece exponencialmente con el número de variables. Entre éstos, el método 'branch and bound' (BB) es uno de los más populares. En él se garantiza encontrar el subconjunto óptimo de un tamaño dado, si la función de la evaluación tiene un comportamiento monotónico. En otras palabras, si un clasificador que utiliza un subconjunto de variables de entrada presenta un éxito de clasificación peor que otro clasificador que utiliza otro subconjunto de variables, se asume que ninguna combinación de las variables presentes en el primer subconjunto conducirá a un mejor éxito en la clasificación y, por lo tanto, debe abandonarse la búsqueda entre esas variables.

Los algoritmos de búsqueda secuenciales siguen estrategias que reducen el número de estados que se analizan durante la búsqueda, aplicando la búsqueda local. Los métodos más comunes son el forward selection (SFS) y el backward selection (SBS). Sin embargo éstos tienden a quedarse atrapados en soluciones sub-óptimas.

Por su parte, los algoritmos de búsqueda aleatorios intentan superar el costo computacional de los métodos exponenciales, Dichas técnicas incluyen los algoritmos genéticos (GA), y el simulated annealing (SA) entre otros [29].

3.5.2 Métodos secuenciales o (deterministas)

Los algoritmos de búsqueda secuenciales son estrategias que reducen el número de variables aplicando búsquedas locales. Los métodos más comunes son la sequential forward selection (SFS) y la sequential backward selection (SBS) [30,31,32,33].

ISBN:978-84-693-4070-7/DL:T-1167-2010

3.5.2.1 Método secuencial forward selection (SFS)

SFS comienza con un conjunto sin variables y secuencialmente va agregando parámetros. El procedimiento continúa hasta que el criterio de selección haya alcanzado un mínimo o todos los parámetros se agreguen al modelo.

El procedimiento empieza considerando cada una de las variables individualmente y seleccionando la variable, (z1) que da el mejor valor obtenido por el criterio de selección (J), donde el criterio de selección generalmente se calcula por medio del error de predicción (PRE) sobre los datos de validación. En nuestro caso se ha implementado dicho criterio mediante el PRE obtenido utilizando redes neuronales (fuzzy ARTMAP y PNN) como clasificadores.

El próximo paso es entonces calcular todos los posibles modelos de dos variables que incluyen a z1 de nuevo. La nueva variable añadida será aquella cuya inclusión reduzca en mayor medida el error de predicción del clasificador neuronal. Este proceso continúa hasta que la inclusión de nuevas variables no reduzca el PRE o bien todas las variables hayan sido incluidas. La figura 3.9 muestra el diagrama de flujo de este método.

3.5.2.2 Método secuencial backward selection (SBS)

El método (SBS) funciona de forma contraria al forward selection. En este caso, todas las variables del conjunto son incluidas al principio para ser utilizadas por el clasificador. Las variables en este caso se van descartando o eliminando en un momento dado basándose en su contribución al criterio de selección J. Es decir, se van eliminando secuencialmente aquellas variables cuya exclusión no degrada el PRE del clasificador. En la figura 3.10 se muestra el diagrama de flujo del proceso secuencial de este método de selección.

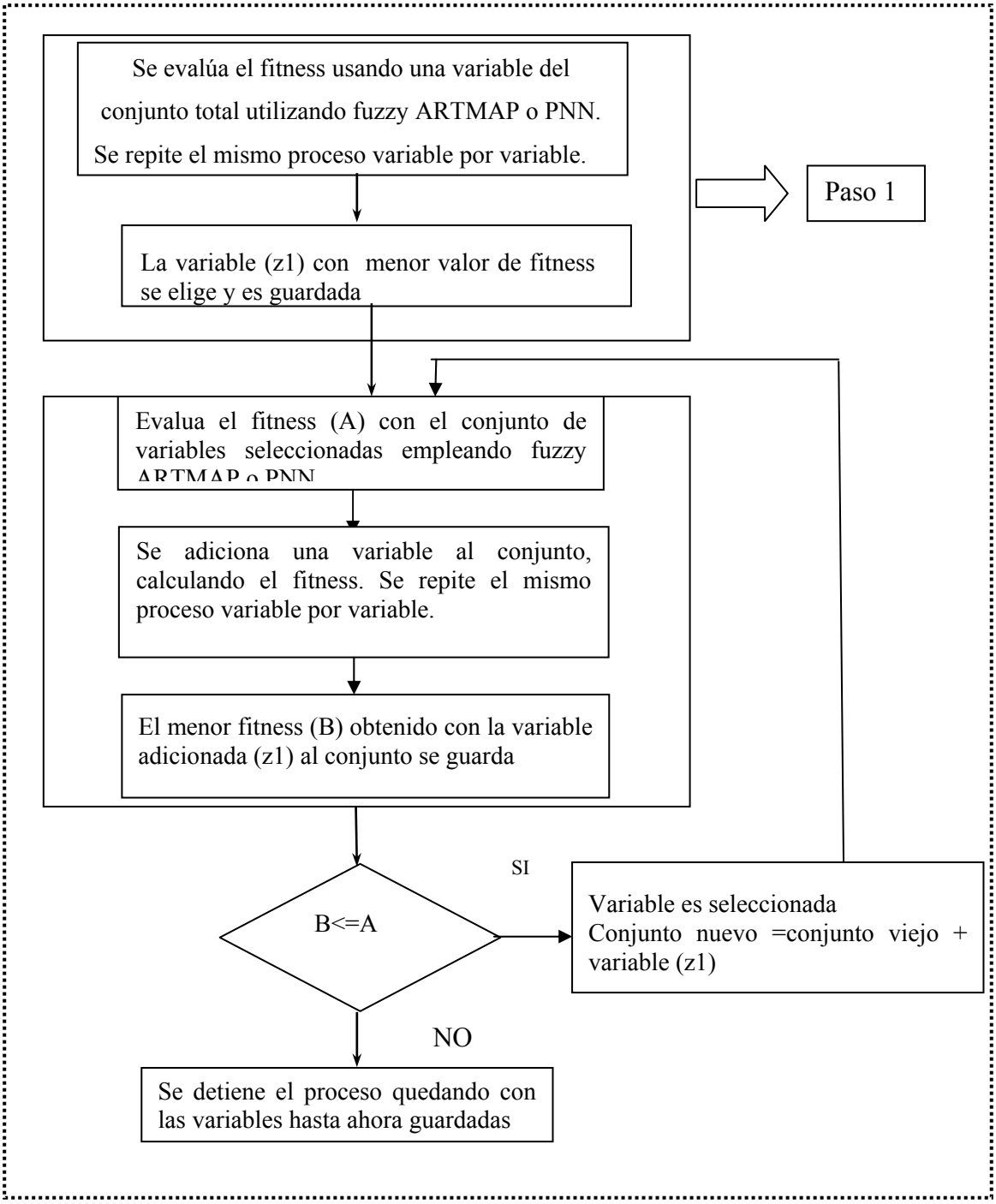

Figura.3. 9 Diagrama de flujo del método secuencial forward selection

ISBN:978-84-693-4070-7/DL:T-1167-2010

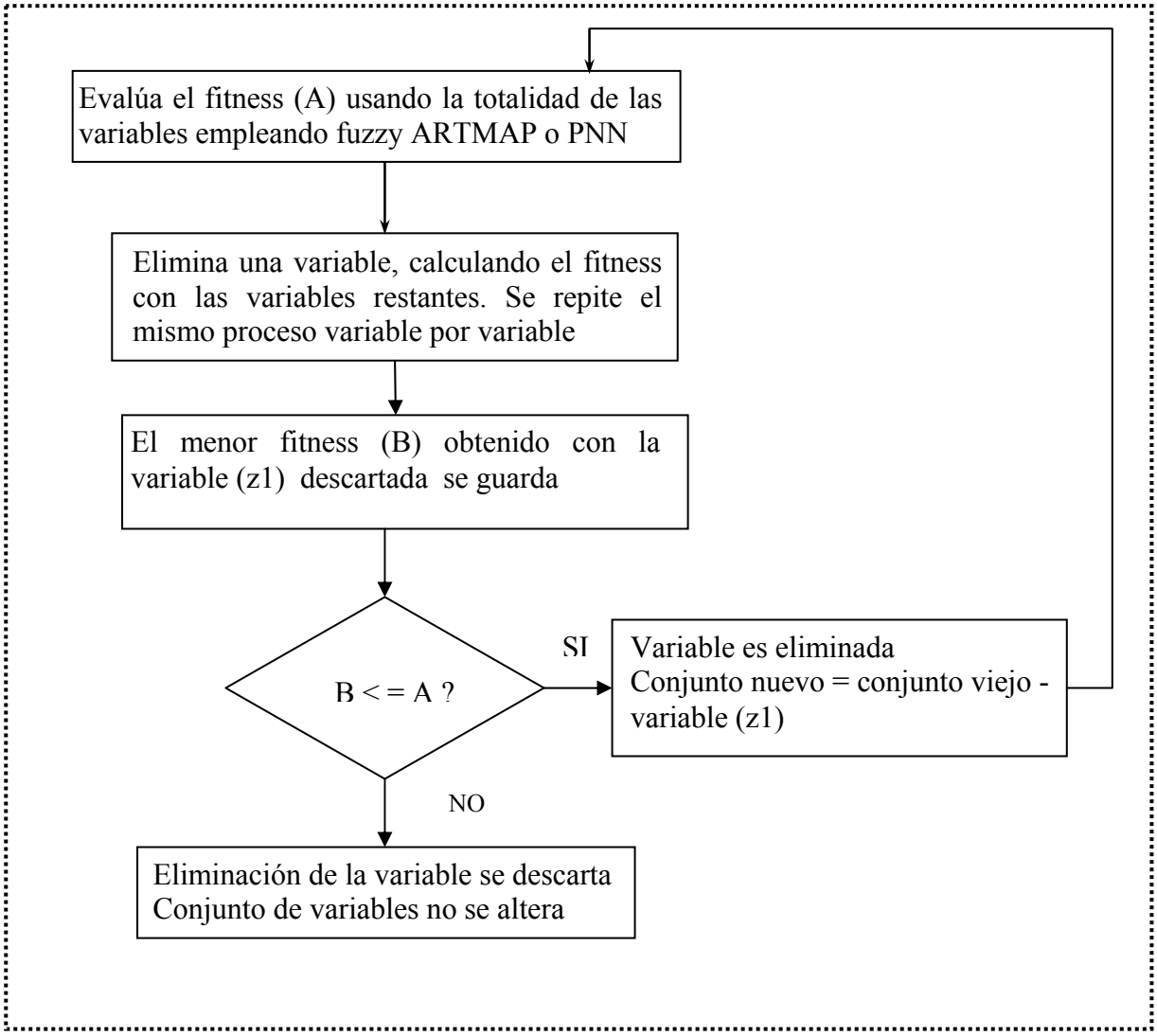

Figura. 3.10 Diagrama de flujo método secuencial backward selection

Existe un método que unifica los anteriores como es el stepwise selection method [2]. Este consiste en el encadenamiento de fases SBS y SFS. De esta manera, una variable que entró (fue eliminada) del modelo en las fases iniciales de selección puede volver a ser incluida en las fases posteriores. Una vez más, el número de variables retenidas en el modelo se basa en que su utilización como variables de entrada contribuya a disminuir el error de predicción del clasificador neuronal.

3.5.3 Métodos de optimización estocásticos

Los métodos no estocásticos revisados en la sección anterior permiten, en muchos casos, optimizar el número de variables que deben emplearse como entradas de un clasificador. Sin embargo, cuando el número de variables crece, dichos métodos tienden

a quedarse atrapados en soluciones subóptimas. En estos casos, puede ser interesante abordar el problema de selección de variables empleando métodos estocásticos.

Estos permiten realizar búsquedas locales alrededor de soluciones prometedoras pero poseen la componente de aleatoriedad que les permite explorar otras soluciones en el espacio de búsqueda. Los dos métodos estocásticos más empleados para abordar problemas de optimización son los algoritmos genéticos y las técnicas de recocido simulado (simulated annealing). A continuación se revisan ambos métodos:

3.5.3.1 Algoritmos genéticos

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 generada aleatoriamente [3, 8, 31, 34,35, 36,37,39].

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 se llama gen y un juego de variables se llama un cromosoma. Por ejemplo, en un problema de selección que empieza con 8 variables, un posible cromosoma sería 00110101. Esto puede traducirse tal que las variables 3, 4, 6 y 8

ISBN:978-84-693-4070-7/DL:T-1167-2010

serán usadas en el proceso de modelado (para entrenar y validar un clasificador neuronal) y las variables 1, 2, 5 y 7 serán omitidas [8].

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 obtenido por redes clasificadoras fuzzy ARTMAP y PNN.

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 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 generación 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 "crossover point", o punto de cruce, es un valor aleatorio. La figura 3.11 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 permitiendo sobrevivir a las combinaciones que mejor cumplen con el criterio preestablecido tras aplicarles combinaciones y mutaciones, permite llegar a una serie de patrones (denominados "schematas") que convergen a la solución óptima del problema.

Entre las características de la implementación particular [9] sobre los algoritmos genéticos que se ha utilizado en esta tesis, basada en Matlab y el PLS Toolbox, destacaremos las siguientes:

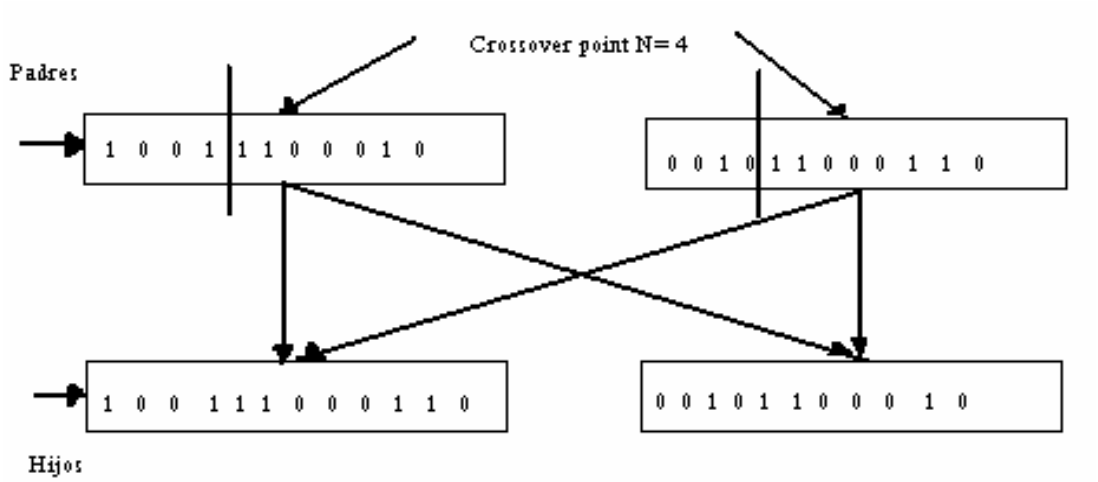

Figura 3.11 esquema explicativo del cruce entre padres para la generación de hijos

- 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 son los siguientes:

- Número de miembros de la población: indica el número de individuos que deben considerarse y evaluarse en cada iteración.
- Número de términos iniciales: número de miembros en la población inicial.
- Número 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.

ISBN:978-84-693-4070-7/DL:T-1167-2010

- Probabilidad de mutación: fracción de bits que deben ser cambiados en cada generación.
- Tipo de combinación: "crossover simple o doble"
- Número de subconjuntos en validación cruzada: como su nombre lo indica, número de subconjuntos que serán iterativamente utilizados para entrenar y evaluar el fitness de cada miembro de la población.
- Número de iteraciones: Número de veces que se deben formar los subconjuntos para obtener un fitness medio con menor "rizado" debido a la aleatoriedad.

3.5.3.2 Algoritmo Simulated Annealing

Simulated annealing es una técnica estocástica de optimización que permite hallar soluciones cercanas al óptimo global en problemas de optimización complejos (p.e. con elevado número de variables) [41-47]. Este método de optimización fue introducido por Kirkpatrick y colaboradores [48], basándose en trabajos presentados por Metrópolis y colaboradores [49]. Esta estrategia simula el proceso de cocción ("annealing") de los metales.

Un annealing (o proceso de recocido) es un proceso en el cual un sólido es fundido aumentando la temperatura a un valor elevado para posteriormente reducirla lentamente hasta que el sistema se enfría adoptando una configuración de energía mínima. [50].

En el problema de selección de variables, el algoritmo progresa a través de un espacio de búsqueda definido por el universo de todas las combinaciones posibles de variables que pueden ser utilizadas para resolver un problema determinado (p.e. identificación o cuantificación de mezclas de gases).

El algoritmo parte de la utilización de un conjunto de variables (generalmente todas las variables disponibles), X_0. Una vez calculado el fitness (error de predicción obtenido con la red clasificadora fuzzy ARTMAP o PNN) asociado a X_0, se determina un conjunto X_1 resultado de haber eliminado *n* variables (escogidas

aleatoriamente dentro de X_0). El fitness de X_1 es obtenido y se calcula la diferencia entre el modelo original y el nuevo.

$$\Delta E = \text{Fitness (nuevo)} - \text{Fitness (viejo)} \quad (3.34)$$

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

$$P_i = \exp(-\Delta E / T_i) \quad (3.35)$$

donde T_i es la temperatura de trabajo (se escoge un valor inicial para T_i). Si $P_i > R$ (R es un valor aleatorio con distribución uniforme entre [0,1]) la nueva solución es retenida y el algoritmo prosigue eliminando variables a partir de X_1. En caso contrario, el algoritmo prosigue desde X_0. La figura 3.13 muestra a *p* como una función de los cambios de fitness (valores positivos de ΔE) a diferentes temperaturas annealing T_i. Esta figura muestra que cuando la temperatura de annealing se reduce, la posibilidad para aceptar una solución peor decrece significativamente.

El proceso se ejecuta un número determinado de iteraciones (# iteraciones = N variables totales - 1) para la temperatura T_i. Finalmente todo lo anterior se repite para $T_{i+1} = \alpha T_i$ ($\alpha < 1$). En otras palabras, el proceso de selección se desarrolla a una temperatura mas baja (el proceso de aceptación de cambios es menos exigente). El número de temperaturas a computar también debe ser definido a priori. Teniendo en cuenta que un número bajo puede dar resultado a soluciones no muy buenas y un número alto incrementaría drásticamente el tiempo de computación del proceso (un valor promedio de ejecución puede ser de 50 temperaturas). En la figura 3.12 se puede ver el diagrama de flujo del algoritmo simulated annealing.

ISBN:978-84-693-4070-7/DL:T-1167-2010

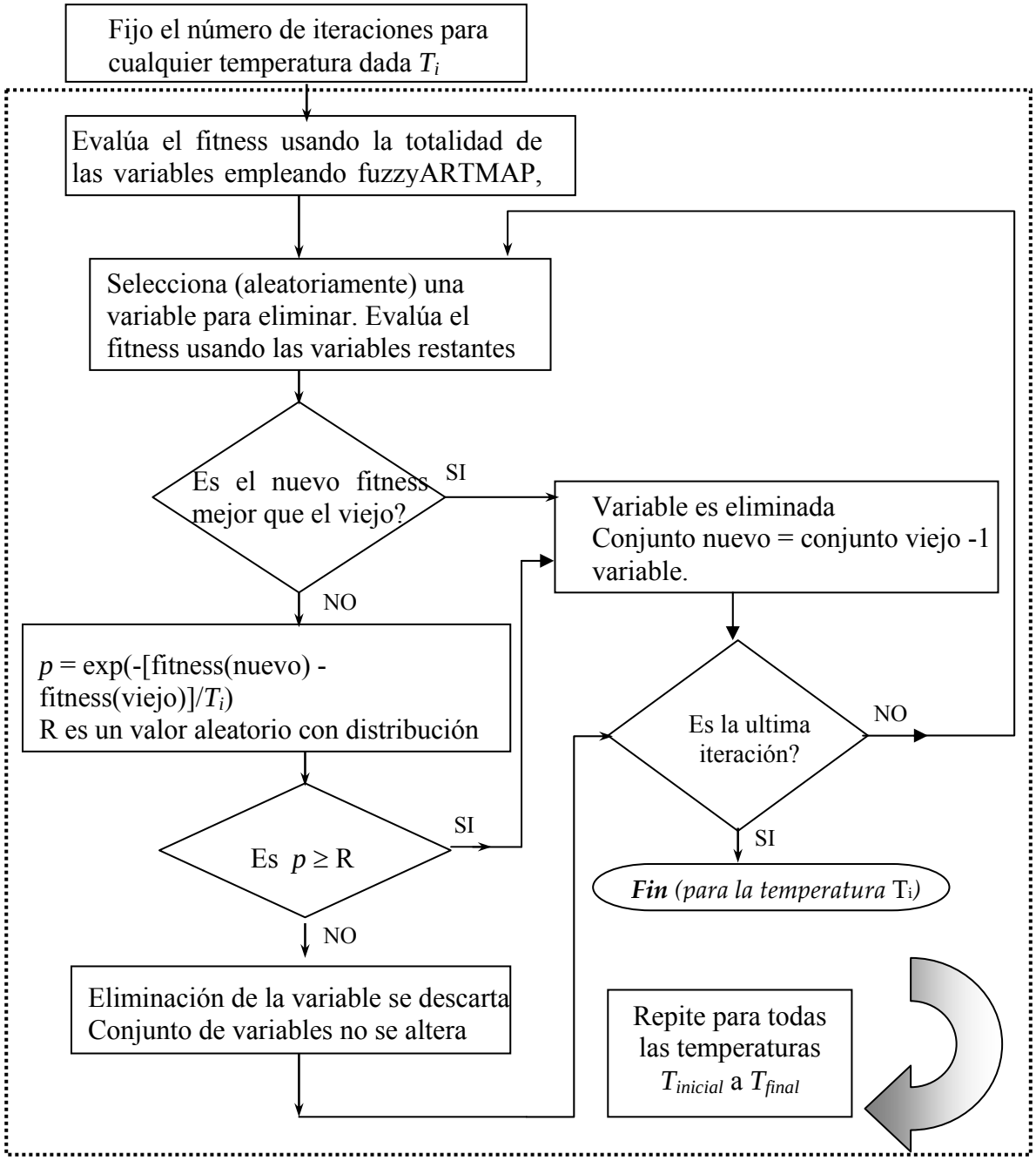

Figura 3.12 Diagrama de flujo del algoritmo simulated annealing implementado para la selección de variables.

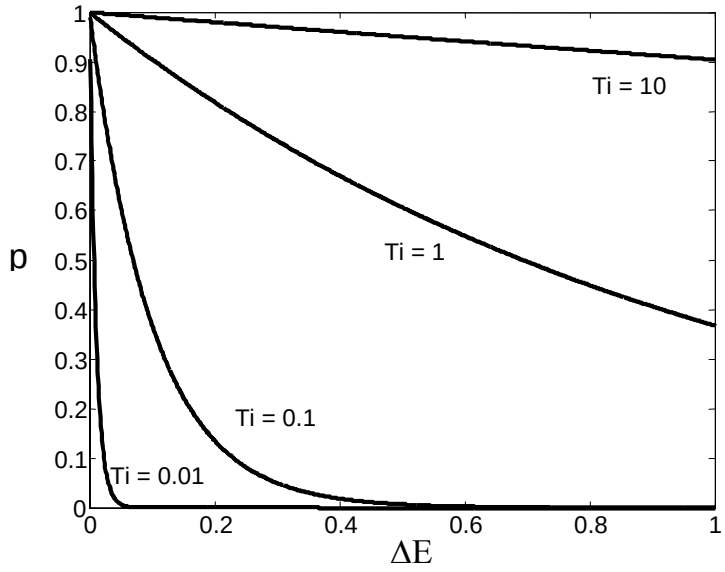

Figura 3.13 probabilidad p de aceptación a soluciones peores como una función del cambio de fitness a diferentes temperaturas annealing, T_i

3.6 Técnicas o estrategias de selección de variables para eliminar variables redundantes, ruidosas y con información irrelevante.

Como se ha mencionado anteriormente, en las aplicaciones con sistemas de olfato electrónico es habitual que se utilicen medidas de gran dimensionalidad (muchas variables por medida que son parámetros extraídos de la respuesta de los sensores). Entre otros, uno de los grandes inconvenientes que se puede encontrar es que al tener un alto número de variables el tiempo de cálculo empleado para la selección de las variables óptimas es, en muchos casos, inaceptablemente elevado aún utilizando estrategias de selección estocásticas o secuenciales.

Por este motivo, tiene sentido utilizar técnicas de "pre-selección de variables" que eliminen del conjunto aquellas variables que sean redundantes o introduzcan ruido dentro del sistema en un tiempo mínimo sin que ello afecte significativamente al

ISBN:978-84-693-4070-7/DL:T-1167-2010

posterior proceso de clasificación. Dicho proceso de pre-seleccion se ha realizado en dos pasos que se mencionan a continuación:

3.6.1 Primer paso: "Criterio de la varianza"

En un primer paso, previo a la selección final de variables, se buscaría implementar un criterio capaz de determinar la habilidad de discriminación de cada variable (por ejemplo, las relaciones m/z en un espectrómetro de masas). Las medidas usadas para la fase de entrenamiento se agrupan en categorías (es decir, medidas de un mismo tipo de gas o de tipos de jamón por ejemplo, se agrupan en una categoría). 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.

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 los 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 (o variables) 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.

La figura 3.14 y las ecuaciones (3.35), (3.36) y (3.37) detallan un cálculo de resolución a modo de ejemplo.

Los pasos detallados de este cálculo son los siguientes:

1. Cálculo del 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).

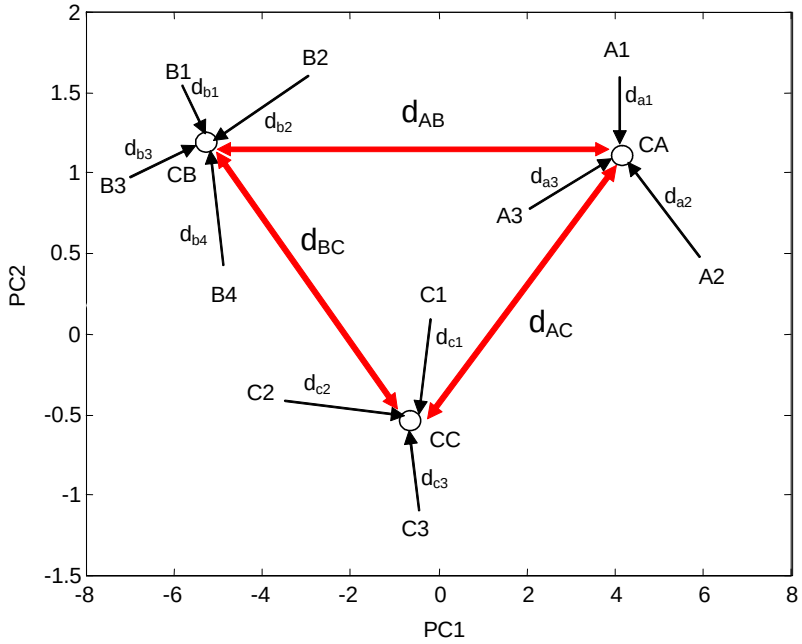

Figura 3.14: Cálculo del poder de resolución con 10 medidas de tres clases diferentes

Para cada clase, la distancia intraclase media es:

$$v_A = \frac{1}{3}(da1 + da2 + da3) \quad v_B = \frac{1}{4}(db1 + db2 + db3 + db4) \quad v_C = \frac{1}{3}(dc1 + dc2 + dc3) \tag{3.35}$$

Distancia intraclase media: $vm = \frac{1}{3}(v_A + v_B + v_C)$ (3.36)

Distancia media interclases: $vmi = \frac{1}{3}(d_{AC} + d_{AB} + d_{BC})$ (3.37)

ISBN:978-84-693-4070-7/DL:T-1167-2010

Poder de resolución (DA_j): $res = \frac{vmi}{vm}$ (3.38)

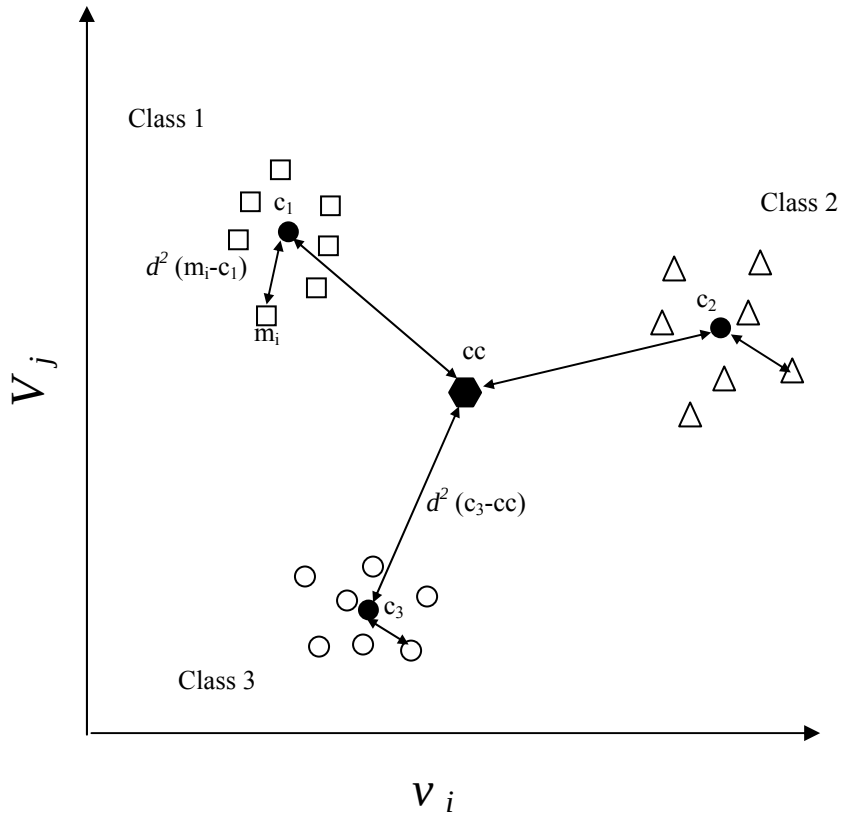

Figura 3.15 *interpretación geométrica del proceso usado para calcular la figura de mérito para dos variables cualesquiera dadas (caso de tres categorías). La varianza intra categoría se calcula como el promedio de las distancias cuadradas entre medidas dentro de una categoría. (por ejemplo m_i) y el centroide de la categoría (por ejemplo c_1). La varianza entre categorías se calcula como el sumatorio de las distancias entre los centroides de las categorías y el centroide total (cc). d^2 (m_i-c_1) denota la distancia cuadrada Euclidea entre la medida i y el centroide de su categoría, c_1.*

Cuanto más alto sea el poder de resolución o la habilidad de discriminación de una variable dada, más importante es esta variable para discriminar correctamente entre las categorías objeto de clasificación. En otras palabras, el ruido o información no deseada para cada variable será asociada a valores bajos de la habilidad de discriminación (DA_j). Por lo tanto, un conjunto de variables, que comprenda aquellas que tengan un valor por encima de un umbral determinado se selecciona para su consideración en posteriores fases de selección de variables. Este método podría ser equivalente al calculo del discriminante lineal de Fisher si el número de categorías para clasificar las medidas fuese de $d = 2$. Este proceso es univariante y puede existir el riesgo de eliminar variables sinérgicas que tienen una habilidad de discriminación baja cuando se consideran individualmente. Para minimizar este problema el proceso descrito anteriormente se repite considerando todas las posibles combinaciones entre 2 variables. La figura 3.15 ilustra este proceso, dando como resultado una nueva lista de figuras de mérito, DA_{ij} es decir, la habilidad para discriminar cuando se usan variables *i* y *j* simultáneamente. Esto permite reseleccionar variables que habían sido descartadas previamente, si se encontrara un efecto sinérgico.

Sin embargo, hay que decir que este método no evita que variables redundantes (es decir altamente colineales) sean seleccionadas durante el proceso. Dicho problema se analizará en un segundo paso de selección que se comentará posteriormente.

3.6.2 Segundo paso: Detección y eliminación de variables redundantes "Colinealidad" entre las variables.

Como segundo paso se hace necesario implementar un proceso que nos permita eliminar variables redundantes (altamente colineales). Partiendo de que R es la matriz (n×p) de calibración resultante del paso anterior, el número de columnas, p corresponde al número de variables seleccionadas en el primer paso, y el número de filas n, corresponde al número de medidas dentro del conjunto de calibración.

ISBN:978-84-693-4070-7/DL:T-1167-2010

$$R^t = \begin{bmatrix} m/z_{1,m1} & m/z_{1,m2} & \cdots & m/z_{1,mp} \\ m/z_{2,m1} & m/z_{2,m2} & \cdots & m/z_{2,mp} \\ \vdots & \vdots & \cdots & \vdots \\ m/z_{n,m1} & m/z_{n,m2} & \cdots & m/z_{n,mp} \end{bmatrix} \tag{3.39}$$

R^t *describe la transpuesta de R* donde $m/z_{i,mj}$ corresponde a la intensidad de las relaciones masa/carga para cada medida *j* (aunque también pueden ser los parámetros extraídos de los sensores). Para cada variable, un vector respuesta normalizado puede ser definido de la siguiente manera:

$$m/z_i = \left(\frac{m/z_{i,m1}}{\sqrt{\sum_{j=1}^{p} m/z_{i,mj}^2}}, \frac{m/z_{i,m2}}{\sqrt{\sum_{j=1}^{p} m/z_{i,mj}^2}}, \cdots, \frac{m/z_{i,mp}}{\sqrt{\sum_{j=1}^{p} m/z_{i,mj}^2}} \right) \text{ for } i = 1 \text{ to } n. \tag{3.40}$$

La ecuación muestra el vector respuesta normalizado para la relación masa/carga *i*. Ahora, el grado de colinealidad existente en el conjunto de calibración entre dos masas diferentes puede ser determinado calculando el producto escalar de sus vectores respuestas normalizados como se muestra a continuación:

$$P_{i,k} = \sum_{q=1}^{p} \left(\frac{m/z_{i,mq}}{\sqrt{\sum_{j=1}^{p} m/z_{i,mj}^2}} \times \frac{m/z_{k,mq}}{\sqrt{\sum_{j=1}^{p} m/z_{k,mj}^2}} \right) \tag{3.41}$$

$P_{i,k}$ es el producto escalar entre los vectores respuesta normalizados asociados a la variable *i* y *k*, donde los posibles valores de $P_{i,k}$ están comprendidos entre 0 y 1. El valor mas cercano a la unidad de $P_{i,k}$ es el que presenta mayor colinealidad entre las variables *i* y *k*. El número de productos escalares esta dado por $\sum_{i=1}^{n-1}(n-i)$.

Una vez obtenidos los productos escalares, se determina un umbral que es usado para decidir que variables son redundantes y deberían ser eliminadas o no tenidas en cuenta. Este segundo paso de selección de variables es no supervisado puesto que no hay necesidad de clasificar las muestras de entrenamiento de acuerdo a su categoría.

Después de haber eliminado aquellas variables con ruido, con información irrelevante y/o redundante, el conjunto de variables supervivientes estará listo para realizar el siguiente y ultimo paso en el proceso de selección (es decir utilizando los diferentes métodos de selección tanto secuenciales como estocásticos). La figura 3.15 muestra un ejemplo del producto escalar usando una base de datos típica de amoníaco, acetona y tolueno

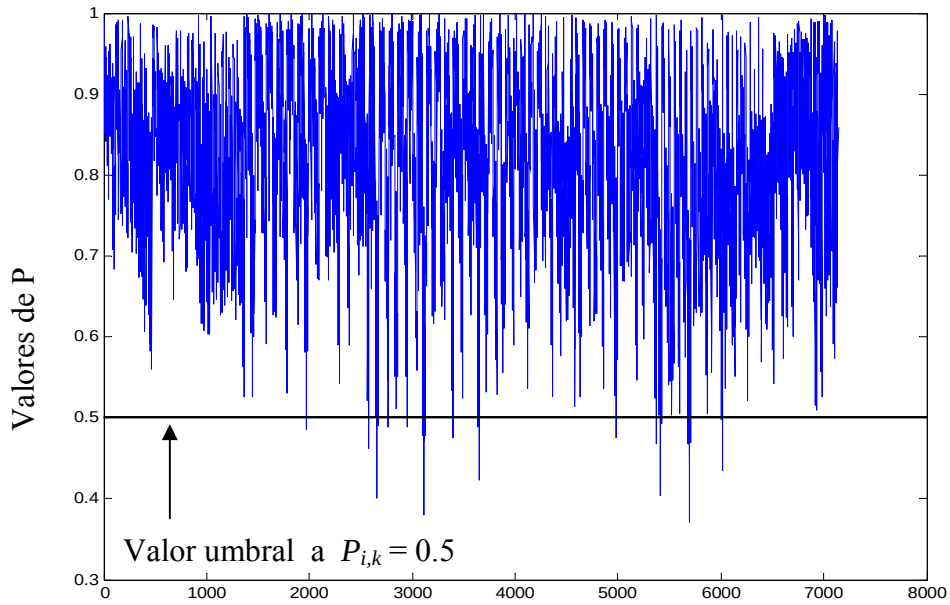

Figura 3.16: Gráfico del producto escalar usando una base de datos de amoníaco, acetona y tolueno. [120 variables proporcionan 7,140 productos escalares calculados]

3.7 Conclusiones

En este capítulo se han presentado las bases teóricas fundamentales relacionadas con los diferentes métodos de selección de variables y técnicas de reconocimiento de patrones que se estudian y evalúan en esta tesis doctoral. Todas ellas se han utilizado en el análisis de las diferentes aplicaciones prácticas tratadas en este trabajo.

El capítulo comienza definiendo una de las partes más importantes en los sistemas de olfato electrónico como son los algoritmos o técnicas de reconocimiento de patrones, mostrando las diferentes formas en las que se pueden clasificar dependiendo de sus características (supervisado o no supervisado, paramétricos o no paramétricos). Una vez hecha la clasificación se describen con minucioso detalle las técnicas utilizadas en este trabajo, las redes fuzzy ART, fuzzy ARTMAP y PNN. Por otro lado, también se define otra familia de algoritmos de reconocimiento de patrones como son los Support Vectors Machines (SVM) utilizados para los diferentes procesos de selección, clasificación y regresión (sección 3.4).

Además, en la sección 3.5 se menciona resumidamente las nociones básicas y las posibles implicaciones de la selección de variables en los sistemas multisensoriales. En esta sección se han incluido los subapartados donde se detallan los diferentes métodos de selección como son los métodos secuenciales (forward, backward y stepwise) y los métodos estocásticos (algoritmos genéticos y simulated annealing). Finalmente en la sección 3.6 se detallan otras técnicas o estrategias de selección de variables utilizadas para eliminar variables redundantes, ruidosas y con información irrelevante como son las técnicas de la varianza y la colinealidad.

3.8 Referencias

[1] Ricardo Gutiérrez "Pattern analysis for machine olfaction: A review" IEEE sensor journal, vol 2 no 3, june 2002

[2] Pearce, T.C. Schiffman, S.S. Nagle,H.T. Gardner, J.W. "Handbook of Machine Olfaction" Electronic nose Technology. Editorial Wiley-Vch. 2003

[3] Julian Gardner and Philip Bartlett "Electronic Noses principles and aplications" edit Oxford Science Publications 1999.

[4] Margalef Pedrola Pere Joan, "Tesis Avaluacio de xarxes ressonants en la clasificacio de mostres gasoses". Universidad Rovira i Virgili . junio 2002.

[5] 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, 1992, Vol. 3, Núm. 5, pàgs. 698-713.

[6] E. Llobet, 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 Actuators B vol 99, 267-272 (2004) .

[7] R. Ionescu and E. Llobet, "Olfato electrónico: Técnicas avanzadas de reconocimiento de patrones para sistemas de nueva generación (Electronic noses: Advanced techniques of pattern recognition for new generation systems) – in Spanish", Mundo Electrónico, vol. 317, pp. 48-52, 2001.

[8] Brezmes Jesus "Tesis Diseño de una nariz electrónica para la determinación no destructiva del grado de maduración de la fruta" Universidad Politécnica de Cataluña 1999.

[9] The Mathworks Inc., Matlab (versió 6.0), The Mathworks. Inc http:// www.mathworks

ISBN:978-84-693-4070-7/DL:T-1167-2010

[10] V.N. Vapnik, Estimation of Dependencies Based on Empirical Data, Springer Verlag, New York, 1982 (first published in Russian, Nauka, Moscow, 1979).

[11] V.N. Vapnik, A.Y. Chervonekis, The necessary and sufficient conditions for the consistency of the method of empirical minimization, Pattern Recognition and Image Analysis, 1 (1991) 284-305.

[12] V.N. Vapnik, The Nature of Statistical Learning Theory, J. Wiley & Sons, Inc., New York, 1998.

[13] D. Meyer, F. Leisch, K. Hornik, The support vector machine under test, Neurocomputing, 55 (2003) 169-186.

[14] S.A. Khalifa, S Maldonado-Bascón, J.W. Gardner, Identification of CO and NO2 using a thermally resistive microsensor and support vector machine, IEE Proceedings, Sci., Meas. Technol., 150 (2003) 11-14.

[15] C. Distante, N. Ancona, P. Siciliano, Support vector machines for olfactory signals recognition, Sensors and Actuators B, 88 (2003) 30-39.

[16] M.L- Frank, M.D. Fulkerson, B.R. Patton, P.K. Dutta, TiO2-based sensor array modeled with nonlinear regression analysis for simultaneously determining CO and O2 concentrations at high temperatures, Sensors and Actuators B, 87 (2002), 471-479.

[17] S Maldonado-Bascón, S.A. Khalifa, F. López-Ferreras, Feature reduction using suport vector machines for binary gas detection, Lecture Notes in Computer Science, 2687 (2003) 798-805.

[18] T. Quian, R. Xu, C.Kwan, B. Linell, R. Young, Toxic vapor classification and concentration estimation for shuttle and international space station, Lecture Notes in Computer Science, 3173 (2004) 543-551.

[19] K. Brudzewski, S. Osowski, T. Markiewicz, Classification of milk by means of an electronic nose and SVM neural network, Sensors and Actuators B, 98 (2004) 291-298.

[20] O. Sadik, W.H. Land, A.K. Wanekaya, M. Uematsu, M.J. Embrechts, L. Wong, D. Leibensperger, A. Volykin, Detection and classification of organophosphate nerve agent simulants using support vector machines with multiarray sensors, J. Chem. Inf. Compt. Sci., 44 (2004) 499-507.

[21] M. Pardo, G. Sberveglieri, Classification of electronic nose data with support vector machines, Sensors and Actuators B, 107 (2005) 730-737.

[22] M. Bicego, Odor classification using similarity-based representation, Sensors and Actuators B, 110 (2005) 225-230.

[23] K. Brudzewski, S. Osowski, T. Markiewicz, J. Ulaczyk, Classification of gasoline with supplement bio-products by means of an electronic nose and SVM neural network, Sensors and Actuators B, 113 (2006) 135-141.

[24] V.N. Vapnik, S. Golowich, A. Smola, Support vector method for function approximation, regression estimation and signal processing, In Advances in Neural Information Processing Systems 9, MIT Press, Cambridge, MA, USA, 1997.

[25] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Min. Knowl. Disc., 2 (1998) 121-167.

[26] V. Kecman, Support Vector Machines –An Introduction, In Support Vector Machines: Theory and Applications, Springer Verlag, Berlin, 2005.

[27] V. Kecman, Learning and soft computing, support vector machines, neural networks and fuzzy logic models, The MIT Press, Cambridge, MA, USA, 2001.

[28] C.W. Hsu, C.J. Lin, A comparison of methods for multiclass support vector machines, IEEE Transactions on Neural Networks, 13 (2002) 415-425.

[29] David Broadhurst, Royston Goodacre, Alun Jones " Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry" Analytica chimica acta 348 (1997) 71-86.

[30] E. Llobet, R. Ionescu, S. Al-Khalifa, J. Brezmes, X. Vilanova, X. Correig, N. Bârsan and J.W. Gardner, "Multicomponent gas mixture analysis using a single tin oxide sensor and dynamic pattern recognition", IEEE Sensors Journal, vol. 1, iss. 3, pp. 207-213, 2001.

[31] Lu Xu, Wen-Jun Zhang, "Comparison of different methods for variable selection", Analytica Chimica Acta 446 (2001) 477-483.

ISBN:978-84-693-4070-7/DL:T-1167-2010

[32] Nils Paulsson, Larson Elisabeth, Winquist Fredik "Extraction and selection of parameters for evaluation of breath alcohol measurement with an electronic nose", Sensors Actuators A 84 (2000) 187-197.

[33] Sutter J. M., Kalivas J.H. "Comparison of forward selection, Backward elimination, and generalized simulated annealing for variable selection" Microchemical journal 47 (1993) 60-66.

[34] Kailing Tang, Tonghua Li. "Combining PLS with GA-PLS for QSAR" Chemometrics and intelligent laboratory systems, 64 (2002) 55-64.

[35] L Davis, The Hanbook of genetic Algorithms. Van NostrandReinhold, New York, 1991.

[36] A.S Barros, D.N Rutledge, Genetic algorithms applied to the selection of principal components, Chemom. Intell. Lab, Sys, 40 (1998) 65-81.

[37] U. Depezynski, V, J Frost,, K, Molt, Genetic Algorithm applied to the selection of factors in principal component regression, Anal Chim Acta 420 (2000) 217-227.

[38] Brezmes, Jesus; Cabre, Pere; Rojo, Sergi. "Discrimination betwen different samples of olive using variable selection techniques and modfields Fuzzy ARTMAP Neural Networks.

[39] J.W. Gardner; P Boilot; E.L. Hines "Enhancing electronic nose performance bby sensor selection using a new integer-based genetic algorithm approach" Sensors and Actuarors Articule en Press (2004).

[40] Tom Artusson; Martin Holmberg, "Wavelet transform of electronic tongue data" Sensors and Actuators B 87 (2002) 379-391.

[41] Gwo-Ching Liao, Ta-Peng Tsao, Application of fuzzy neural networks and artificial intelligence for load forecasting, Elec. Power Sys. Res., 70 (2004) 237-2444.

[42] Matthew Cambell, Amon, Cristina. Cagan, Jhonatan. "Electronic Component Placement using Simulated Annealing Under thermal Constraint" Departament de engineering design research center. University Pitsburg.

[43] R. Meiri. Jacob, Zahav. "Using simulated annealing to optimize the feature selection problem in marketing applications" European Journal of Operational Research 171 (2006) 842–858.

[44] Alex, Alexandridis. Panagiotis, Patrinos. Haralambos, Sarimveis. George, Tsekouras. "A two-stage evolutionary algorithm for variable selection in the development of RBF neural network models" Chemometrics and Intelligent Laboratory Systems 75 (2005) 149– 162.

[45] J.H. Kalivas, Optimization using variations of simulated annealing, Chemometr. Intell. Lab. Syst. 15 (1992) 1 – 12.

[46] U. Horchner, J.H. Kalivas, Further investigation on a comparative study of simulated annealing and genetic algorithm for wavelength selection, Anal. Chim. Acta 311 (1995) 1– 13.

[47] M. Kudo, J. Sklansky, Comparison of algorithms that select features for pattern classifiers, Pattern Recogn. 33 (2000) 25– 41.

[48] S. Kirkpatrick, C.D. Gelati, M.P. Vechi, Optimization by simulated annealing, Science 20 (1983) 671– 680.

[49] M. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, E. Teller, Equation of state calculations by fast computing machines, J. Chem. Phys. 21 (1953) 1087–1092.

[50] L. Nolle, D.A. Armstrong, A.A. Hopgood, J. A. Wware, Simulated annealing and genetic algorithms applied to finishing mill optimisation for hot rolling of wide steel strip, Int. J. of Know.-based Intell. Engin. Sys., 6, (2002) 104-111.

UNIVERSITAT ROVIRA I VIRGILI
DESARROLLO DE DIFERENTES MÉTODOS DE SELECCIÓN
DE VARIABLES PARA SISTEMAS MULTISENTORIALES
Oscar Eduardo Gualdron Guerrero
ISBN:978-84-693-4070-7/DL:T-1167-2010

UNIVERSITAT ROVIRA I VIRGILI
DESARROLLO DE DIFERENTES MÉTODOS DE SELECCIÓN
DE VARIABLES PARA SISTEMAS MULTISENTORIALES
Oscar Eduardo Gualdron Guerrero
ISBN:978-84-693-4070-7/DL:T-1167-2010

4.

Resultados

ISBN:978-84-693-4070-7/DL:T-1167-2010

4.1 Introducción

En este capítulo se estudia la importancia de los diferentes métodos de selección de variables mencionados en los capítulos anteriores, comprobando si estos métodos ayudan realmente a solucionar el problema de la alta dimensionalidad de las variables presente en la mayoría de los sistemas multisensoriales. Para ello se han escogido y diseñado cuatro aplicaciones reales que nos permiten evaluar y comparar los diferentes métodos implementados.

Cabe notar que todo este trabajo se ha realizado en tres grandes estudios, el primero consiste en aplicar todos los métodos de selección de variables a un conjunto de datos proveniente de unas medidas realizadas con una matriz de sensores de gases a 3 compuestos volátiles diferentes; el segundo estudio consiste en aplicar los mismos métodos a tres aplicaciones diferentes de SDOE basados en espectrometría de masas; finalmente, el tercer estudio se encarga de aplicar el método SVM a los conjuntos de datos utilizados en los dos estudios anteriores.

4.2 Métodos de selección de variables para Sistemas de Olfato Electrónico (SDOE) basados en sensores de gases.

Tal y como se ha descrito en capítulos anteriores de esta memoria, las muestras analizadas mediante una matriz de sensores químicos se suelen describir mediante decenas o centenares de variables (parámetros estáticos, dinámicos, etc). Por este motivo se hace necesario realizar búsquedas sistemáticas que permitan seleccionar un subconjunto óptimo de variables que permitan optimizar los resultados obtenidos mediante la utilización de métodos de reconocimiento de patrones. Aquí presentamos los resultados obtenidos utilizando diferentes métodos de selección acoplados a clasificadores fuzzy ARTMAP y PNN, paradigmas que convergen mucho más rápido con el conjunto óptimo de variables seleccionadas para una aplicación dada. La metodología aplicada se muestra usando un conjunto de datos que incluye medidas de vapores de acetona, amoníaco y ortoxileno y sus respectivas mezclas binarias.

ISBN:978-84-693-4070-7/DL:T-1167-2010

4.2.1 Equipo de medida

El sistema está formado por un conjunto de 12 sensores de gases comerciales (Figaro engineering Inc.) basados en óxidos metálicos (sensores tipo TGS 825(x2), 826, 822(x2), 800(x2), 882(x2), 2160, 2611 y 2620). Los sensores se encuentran en una cámara controlada en temperatura (30ºC) y con una humedad relativa aproximada del 10 %. La cantidad de compuestos volátiles necesarios para crear la concentración deseada en la cámara de sensores se introduce en fase líquida usando jeringas cromatográficas de alta precisión. Un ventilador homogeniza las mezclas dentro de la cámara. Para limpiar la cámara entre medidas se usa aire sintético. Una descripción más detallada del experimento puede encontrarse en [1]. La ejecución de las medidas, almacenamiento y procesado de los datos se hacen a través de un PC.

Para realizar las diferentes medidas se utiliza una cámara construida en el laboratorio de la universidad con dimensiones de 22 x 14 x 16 cm, con lo que con estas dimensiones se dispone de 5 litros de volumen de gas aproximadamente en su interior. En esta cámara se encuentran los siguientes elementos:

- Matriz de sensores
- Ventilador para homogenizar la mezcla.
- Termómetro para el control de la temperatura.
- Higrómetro para el control de la humedad.

La matriz de sensores se sitúa horizontalmente en la parte superior de la cámara, (figura 4.1) fijada a la tapa de tal forma que los sensores queden boca abajo. El ventilador también se fija a la tapa, pero en forma vertical. Para proteger los sensores del flujo directo del aire producido por el ventilador se interponen entre este y la matriz de sensores unas placas metálicas. El ventilador siempre está funcionando, antes y después de realizar cada medida.

Pensando en su limpieza, a la cámara se le ha dotado de una entrada para introducir el aire seco después de cada medida. Dispone también de un pequeño orificio de salida, para la salida del aire seco mientras se esta limpiando. Una vez

terminada la limpieza, se tapan las entradas y salidas de aire para que la cámara quede hermética y lista para realizar otra medida.

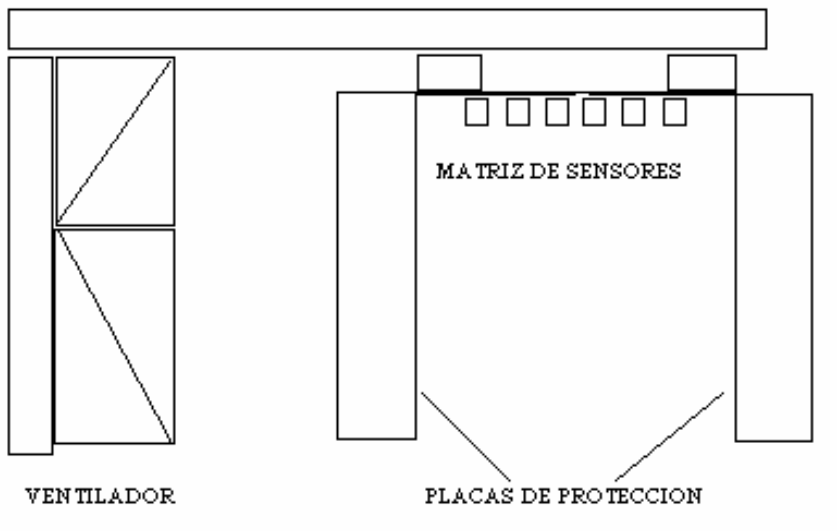

Figura 4.1 Posición de los sensores en la cámara.

También dispone de un pequeño orificio para que pase el cable que conecta la matriz de sensores con la placa interfase. En un costado de la tapa de la cámara se encuentra la entrada por la que se inyecta la mezcla gaseosa. Es un pequeño orificio tapado con septum para introducir una jeringa con la cantidad de mezcla designada. Para esto se utiliza una jeringa de líquidos de una precisión de una décima de microlitro.

La matriz de sensores se conecta a través de un cable plano a una placa interfaz. Esta placa esta conectada a la placa de adquisición de datos instalada en el PC, que permitirán recolectar las variaciones de cada sensor.

ISBN:978-84-693-4070-7/DL:T-1167-2010

4.2.2 Procedimiento de adquisición de las medidas

El proceso de forma detallada que se ha seguido para la adquisición de las medidas es el siguiente:

Partimos de una cámara limpia, unos sensores estabilizados (ventilador siempre encendido) y una humedad controlada del 10 %:

- Se carga la jeringa con el volumen de líquido contaminante apropiado para la concentración que se quiere generar.
- Se inicia el programa de adquisición de datos introduciendo todos los datos sobre el experimento que gestionará el programa.
- Se comienza a adquirir los datos antes de inyectar el contaminante en la cámara. Con este proceso se monitorizará el estado inicial en que se encuentran los sensores antes del experimento, y el cambio brusco presente cuando se realiza la inyección del contaminante. Una alarma sonora nos indicará el momento exacto en el que se debe realizar la inyección.
- Al finalizar la inyección del contaminante se espera 10 minutos para asegurarnos de que los sensores estén totalmente estabilizados.
- Una vez finalizada la adquisición de los datos se realiza la limpieza de la cámara. Se abre la válvula de salida y se inyecta un potente flujo de aire sintético durante dos minutos limpiando así cualquier resto de contaminante y controlando su humedad al 10 %.
- Cuando finalizan los dos minutos volvemos a tapar la salida, y dejamos que los sensores se estabilicen durante una hora como mínimo. Una vez pasado ese intervalo de tiempo, la cámara esta lista para la realización de la siguiente medida.

4.2.3 Conjunto de medidas experimental.

Los vapores medidos fueron acetona, amoníaco y ortoxileno a 50, 100, 200 y 400 ppm y sus mezclas binarias (midiendo todas las posibles combinaciones de dos vapores, de las cuatro concentraciones utilizadas). Estas dan un total de 12 medidas de vapores simples y 48 medidas de mezclas binarias, donde de cada medida se han realizado cuatro repeticiones, dando un total de 240 medidas que conforman el conjunto de datos principal. Cabe recordar que la concentración de un determinado gas será la relación entre el volumen de este gas y el volumen total de la cámara. Multiplicando este resultado por un millón se obtienen las concentraciones en ppm.

MUESTRA	GAS	CONCENTRACION (ppm)
1	Acetona	50
2	Acetona	100
3	Acetona	200
4	Acetona	400
5	Amoníaco	50
6	Amoníaco	100
7	Amoníaco	200
8	Amoníaco	400
9	Ortoxileno	50
10	Ortoxileno	100
11	Ortoxileno	200
12	Ortoxileno	400

Tabla 4.1 Medidas de 12 muestras simples

Las tablas 4.1 y 4.2 muestran las medidas para cada uno de los compuestos volátiles (Acetona, amoníaco y ortoxileno) con sus respectivas concentraciones (50, 100, 200 y 400 ppm) tanto para los vapores simples como para sus mezclas binarias.

MUESTRA	GAS1	CONCENTRACION (ppm)	GAS2	CONCENTRACION (ppm)
1	Acetona	50	Amoníaco	50
2	Acetona	50	Amoníaco	100
3	Acetona	50	Amoníaco	200
4	Acetona	50	Amoníaco	400
5	Acetona	50	Ortoxileno	50
6	Acetona	50	Ortoxileno	100
7	Acetona	50	Ortoxileno	200
8	Acetona	50	Ortoxileno	400
9	Acetona	100	Amoníaco	50
10	Acetona	100	Amoníaco	100
11	Acetona	100	Amoníaco	200
12	Acetona	100	Amoníaco	400
13	Acetona	100	Ortoxileno	50
14	Acetona	100	Ortoxileno	100
15	Acetona	100	Ortoxileno	200
16	Acetona	100	Ortoxileno	400
17	Acetona	200	Amoníaco	50
18	Acetona	200	Amoníaco	100
19	Acetona	200	Amoníaco	200
20	Acetona	200	Amoníaco	400
21	Acetona	200	Ortoxileno	50
22	Acetona	200	Ortoxileno	100
23	Acetona	200	Ortoxileno	200
24	Acetona	200	Ortoxileno	400
25	Acetona	400	Amoníaco	50
26	Acetona	400	Amoníaco	100

Tabla 4.2.a Medidas de 48 muestras binarias

MUESTRA	GAS1	CONCENTRACION (ppm)	GAS2	CONCENTRACION (ppm)
27	Acetona	400	Amoníaco	200
28	Acetona	400	Amoníaco	400
29	Acetona	400	Ortoxileno	50
30	Acetona	400	Ortoxileno	100
31	Acetona	400	Ortoxileno	200
32	Acetona	400	Ortoxileno	400
33	Amoníaco	50	Ortoxileno	50
34	Amoníaco	50	Ortoxileno	100
35	Amoníaco	50	Ortoxileno	200
36	Amoníaco	50	Ortoxileno	400
37	Amoníaco	100	Ortoxileno	50
38	Amoníaco	100	Ortoxileno	100
39	Amoníaco	100	Ortoxileno	200
40	Amoníaco	100	Ortoxileno	400
41	Amoníaco	200	Ortoxileno	50
42	Amoníaco	200	Ortoxileno	100
43	Amoníaco	200	Ortoxileno	200
44	Amoníaco	200	Ortoxileno	400
45	Amoníaco	400	Ortoxileno	50
46	Amoníaco	400	Ortoxileno	100
47	Amoníaco	400	Ortoxileno	200
48	Amoníaco	400	Ortoxileno	400

Tabla 4.2.b (Continuación) Medidas de 48 muestras binarias

La figura 4.2 muestra la respuesta típica de un sensor TGS al ortoxileno y algunos de los parámetros extraídos de la respuesta. La lista completa de parámetros extraídos de las respuestas de los 12 sensores es:

ISBN:978-84-693-4070-7/DL:T-1167-2010

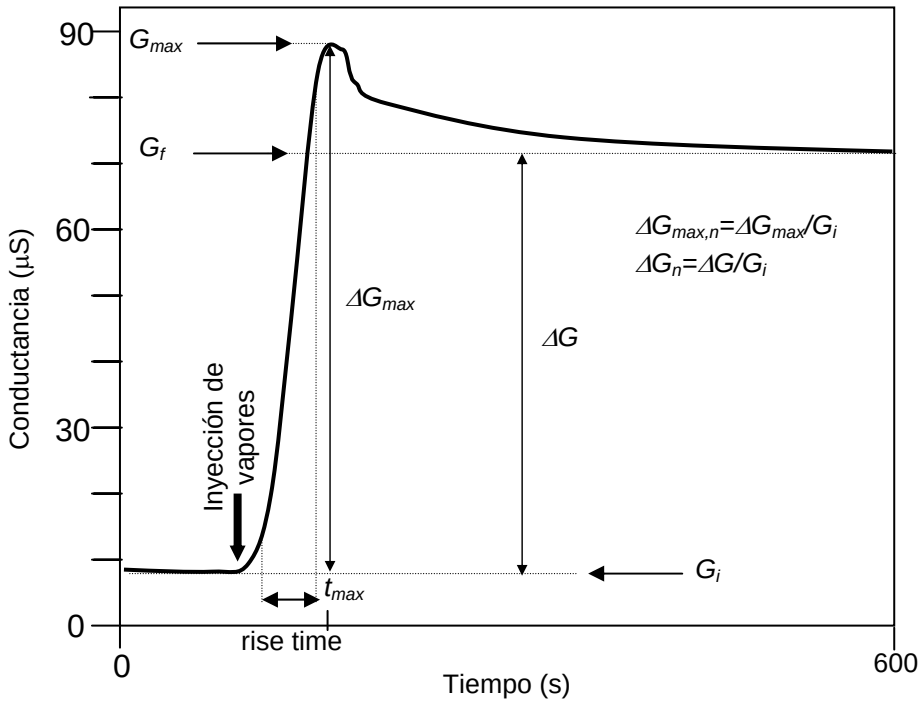

Figura 4.2. Respuesta *típica de un sensor TGS 2610 a 200 ppm de orto-xileno, algunos de los parámetros extraídos de la respuesta del sensor se muestran en la figura*.

- Conductancia inicial (o baseline) G_i .conductancia de un sensor ante la presencia de aire.
- Conductancia final, G_f. conductancia del sensor a el final del periodo de adquisición (600s).
- Incremento de conductancia, $\Delta G= G_f - G_i$.
- Incremento de conductancia normalizada, $\Delta G_n= (G_f - G_i)/G_i$
- Máxima conductancia, G_{max} valor máximo de la conductancia de los sensores (ver figura 1).
- Máximo incremento de conductancia, $\Delta G_{max}= G_{max}- G_i$.
- Máximo incremento de conductancia normalizada, $\Delta G_{max\_n}= (G_{max} - G_i)/G_i$.
- Tiempo en el cual se alcanza la máxima conductancia, t_{max}.

- Tiempo de subida de la conductancia causado por la inyección de compuestos volátiles, definido entre 10% y 90% del máximo cambio de conductancia, t_{10-90}.
- Tiempo de subida de la conductancia causado por la inyección de compuestos volátiles, definido entre 30% y 60% del máximo cambio de conductancia, t_{30-60}.

Estos 10 parámetros se extraen de la respuesta de cada sensor, y debido a que el conjunto de sensores que conforman la matriz es de 12, cada medida se describe con un total de 120 parámetros. La matriz de datos está conformada por 240 filas (medidas) y 120 columnas (variables).

Cada uno de los parámetros se normaliza con respecto a su máximo valor, de forma que el valor de todos los datos presentados a las redes neuronales estarán entre 0 y 1.

4.2.4 Software

Los diferentes métodos de selección de variables se inspiran en rutinas estándard del PLS-toolbox [2]. Dichos métodos, acoplados con redes neuronales fuzzy ARTMAP y PNN fueron programados usando rutinas de MATLAB.

El estudio de este conjunto de medidas se centra inicialmente sobre los vapores simples, realizando los procesos de selección de variables con los diferentes métodos implementados y después se pasará a analizar un conjunto más complejo formado por los vapores simples con sus respectivas mezclas binarias [3,4].

4.2.5 Identificación y cuantificación simultánea de vapores simples

El conjunto de datos inicial es de 48 medidas (4 repeticiones de 3 vapores a 4 concentraciones diferentes). Estos datos se dividen en dos matrices, de forma que cada una contiene 2 medidas con concentraciones idénticas. La primera matriz

ISBN:978-84-693-4070-7/DL:T-1167-2010

(matriz de selección) se usó para aplicar el procedimiento de selección de variables y la segunda (matriz de validación, con medidas diferentes e independiente de la primera matriz) se usó para realizar la validación de los resultados. La matriz de selección se normalizó por columnas. En este procedimiento, cada elemento en la matriz se divide por el máximo valor de su columna. Esta normalización asegura que cualquier elemento en la matriz de selección se encuentre entre [0 1], dándole, de esta forma la misma importancia numérica a las diferentes variables. Este es un aspecto importante, dada la naturaleza heterogénea de los parámetros extraídos (e.g, conductancia, conductancia normalizada, tiempo de subida, etc.). La matriz de validación también se normaliza utilizando el valor máximo de cada una de las columnas calculadas de la matriz de selección.

A partir de este primer pre-procesado se construye un primer modelo fuzzy ARTMAP y PNN utilizando las 24 medidas de la matriz de selección. El número de entradas en el conjunto es de 120 porque este es el número de parámetros extraídos de la matriz de 12 sensores. El número de neuronas a la salida es de 12 porque este es el número de categorías (correspondiente a 3 vapores y 4 concentraciones diferentes). Una vez ajustado el modelo de red (el algoritmo fuzzy ARTMAP aprende a clasificar el conjunto de entrenamiento con un 100% de éxito en sólo una iteración si se utiliza un "learning rate" igual a la unidad), en el proceso de validación las redes se entrenan con la matriz de selección y se evalúan usando la matriz de validación como conjunto de medidas nuevas. La tasa de éxito en la clasificación es de un 41.67 % para ambas redes (es decir 10 de las 24 medidas nuevas fueron identificadas correctamente), claramente una alta tasa de error de predicción de los tipos de vapores y sus concentraciones. Por consiguiente, la utilización de un proceso de selección de variables queda plenamente justificado con el fin de ver si dichos resultados pueden mejorarse.

En la tabla 4.3 se puede observar que acoplando las redes neuronales (fuzzy ARTMAP y PNN) con los diferentes métodos de selección de variables desarrollados utilizando la matriz de selección se pueden obtener buenos resultados en la identificación de los vapores simples reduciendo el número de variables de forma

considerable e incrementando los porcentajes de acierto en la validación de las medidas con las variables seleccionadas. En la tabla se observa que usando el forward selection se obtienen un total de 3 variables seleccionadas empleando la red fuzzy ARTMAP y 6 a través de la red PNN con unos porcentajes de acierto de 79.16 y 83.33 % respectivamente tras realizar el proceso de validación con las variables seleccionadas. Este proceso se realiza de forma similar al hecho con la totalidad de las variables es decir utilizando la matriz de selección con las variables seleccionadas para el entrenamiento de las redes y evaluándolas usando la matriz de validación como conjunto de medidas nuevas. Es importante mencionar que este proceso se realiza a todos los resultado obtenidos con los diferentes métodos de selección.

Los porcentajes de aciertos obtenidos con los parámetros seleccionados por medio del método backward elimination acoplados con los modelos son de 87.5 y 91.66 % respectivamente. En este caso, el número de parámetros seleccionados son 7 con la fuzzy ARTMAP y 32 con la PNN. Cabe recordar que este método funciona de forma dual al Forward Selection. En este caso, todas las variables del conjunto se incluyen inicialmente, para luego irse descartando o eliminando dependiendo del criterio de selección (p.e el error de predicción o PRE).

De la tabla 4.3 se puede deducir que los resultados obtenidos usando los 2 métodos secuenciales clásicos (forward and backward selection) no son muy buenos. Esto puede ser explicado por el hecho de que una vez la variable ha entrado en el modelo, no puede quitarse en el método forward. Debido a esto, uno nunca estará seguro de la verdadera optimización del subconjunto de variables. Igualmente, en el backward elimination, una vez la variable se ha eliminado del modelo no puede ser incluida de nuevo. Debido a esto, tanto la fiabilidad del subconjunto obtenido como la obtención de combinación óptima queda en duda en ambos métodos.

El método stepwise, como se mencionó anteriormente, es esencialmente un backward elimination, aunque la variable eliminada puede seleccionarse de nuevo por medio de un forward. En suma, es la combinación en cierta forma de los dos métodos anteriores. De la tabla se puede observar que el número de variables

ISBN:978-84-693-4070-7/DL:T-1167-2010

seleccionadas con el método stepwise es de 10 y 34, básicamente son las mismas seleccionadas con la backward y las obtenidas en el segundo ciclo con la forward. Los porcentajes de acierto son 95.83% con el modelo fuzzy ARTMAP y 95.83 % con el modelo PNN. Por lo tanto este método lleva a mejores resultados que los anteriores, lo que prueba que la posibilidad de "arrepentirse" de haber añadido o eliminado una variable es una ventaja del stepwise selection respecto a los dos métodos anteriores Además, dichos porcentajes de acierto son mejores que los que se habían obtenido sin realizar selección de variables.

En el caso de la aplicación de algoritmos genéticos, la población inicial (conjunto inicial de posibles soluciones) está formada por 256 cromosomas generados aleatoriamente. El tamaño de la población se mantiene de generación en generación. El número promedio de genes con 1 dentro de la población inicial de cromosomas fue del 50% (término inicial). En cada una de las generaciones, la estructura del modelo representado por la población se valida mediante validación cruzada. Cada miembro de la población se ordena de acuerdo a su fitness. El fitness, en este caso, representa el error de predicción (PRE) resultante de la validación cruzada del modelo fuzzy ARTMAP o PNN construido. Si el resultado del PRE es cada vez mas bajo, entonces el fitness es cada vez mejor, (un error de predicción muy bajo quiere decir que el resultado del entrenamiento con el modelo fue muy bueno). En cada generación, a la mitad de los cromosomas con el error de predicción más bajo se les permite sobrevivir y engendrar una nueva generación. Pares de estos cromosomas son seleccionados aleatoriamente para engendrar nuevas generaciones usando técnicas de cruce de doble punto. También, algunos de los genes en los cromosomas se cambian al azar (la razón de mutación fue de 0.005) después de cada generación. Los cromosomas con mejor fitness son guardados sin alterarlos en la siguiente generación (elitismo). Se realizan iteraciones hasta cuando la población converge o hasta cuando se llega al número máximo prefijado de iteraciones (250). La población converge cuando el porcentaje de cromosomas duplicados en la población es alto (por ejemplo del 80%).

En un experimento adicional, en esfuerzo por evitar la selección de variables irrelevantes, se implementó una serie de GA en cascada. El proceso es el siguiente:

Inicialmente se aplicó un GA acoplado a un clasificador fuzzy ARTMAP o PNN para seleccionar entre las 120 variables extraídas de la respuesta de los sensores. Se usó una validación cruzada para determinar el fitness (error de predicción obtenido con la matriz de selección) de cada cromosoma dentro de la población. El algoritmo converge después de 124 iteraciones. Esta convergencia implica que en la generación 124, el 80% de los cromosomas son idénticos. Para cromosomas con un PRE muy parecido, aquel que haga uso de menos variables es seleccionado. Por consiguiente, 23 de las 120 variables iniciales se seleccionan (PRE de 0.208).

Un segundo GA se aplicó a las 23 variables seleccionadas en el primer GA. En este caso, el número de variables consideradas para la selección es mucho menor que en el primer GA (120). La complejidad de el problema es más baja y el tamaño de la población inicial podría reducirse (i.e 128 en vez de 256). Esto resulta en una operación más rápida del GA. Consecuentemente, el porcentaje de cromosomas idénticos considerando la ocurrencia de convergencia podría incrementarse a un 90% sin aumentar el tiempo de cálculo. El algoritmo converge después de 75 iteraciones. Una vez más, para PRE similares, el modelo que retiene menos variables es el preferido. Después de esta segunda selección, 9 de las 23 variables fueron seleccionadas (con un PRE de 0.042). El factor PRE en el segundo GA es inferior al primer GA, lo que muestra que los algoritmos genéticos en cascada pueden ayudar a eliminar variables irrelevantes que inicialmente fueron seleccionadas en el primer GA. Se comprobó que la aplicación de otro GA no reducía significativamente el PRE, ni decrementaba el número de variables seleccionadas.

Las variables seleccionadas fueron el cambio de conductancia de los sensores 825, 822, 800, 882 y 2620, el máximo cambio de conductancia de los sensores 800 y 2620 y el cambio de conductancia normalizada de los sensores 825 y 2611. Estas 9 variables se utilizaron para construir el modelo fuzzy ARTMAP de la matriz de selección. El porcentaje de clasificación resultante de la validación fue de un 91.67%, resultado muy favorable al compararlo con el resultado obtenido utilizando las 120 variables (41.67%). Los errores consistieron de una muestra de ortoxileno 50

ISBN:978-84-693-4070-7/DL:T-1167-2010

ppm que fue identificada como ortoxileno de 100 ppm, y una muestra de ortoxileno 400 ppm que fue identificada como ortoxileno de 200 ppm. Por consiguiente, los errores provienen de la cuantificación de las muestras, pero desde el punto de vista de la identificación los vapores pueden ser identificados con un 100% de exactitud.

Método de selección normalizado por columnas	# variables seleccionadas	% de aciertos (utilizando la matriz de validación)
Todas las variables Fuzzy ARTMAP	120	41.67%
GA -fuzzyARTMAP	23	91.67%
GA- fuzzyARTMAP 2 cascada	9	91.67%
Forward-fuzzy ARTMAP	3	79.16%
Stepwise-fuzzy ARTMAP	10	95.83%
Backward-fuzzy ARTMAP	7	87.5%
Todas las variables PNN	120	41.67%
GA- PNN	40	91.67%
GA -PNN 2 cascada	7	87.5%
Forward- PNN	6	83.33%
Backward-PNN	32	91.66%
Stepwise-PNN	34	95.83%

Tabla 4.3 Resultados de la validación utilizando los diferentes métodos de selección de variables para el conjunto de medidas de vapores simples.

4.2.6 Identificación de vapores simples y sus mezclas binarias

El conjunto base con el cual se realizó la selección de variables esta formado por 96 medidas (4 repeticiones de las medidas de los tres vapores a 4 concentraciones diferentes y sus mezclas binarias a igual concentración, es decir 48 medidas corresponden a vapores simples y 48 medidas mas a sus mezclas binarias). Este

conjunto fue dividido en dos matrices de 72 y 24 filas respectivamente. La primera matriz contiene tres medidas replicadas de cada tipo, mientras que la segunda sólo contiene una repetición. El primer subconjunto (matriz de selección) se utilizó para realizar el proceso de selección de las variables y el segundo subconjunto (matriz de validación) fue empleado como conjunto de medidas nuevas en el proceso de validación. Dichos subconjuntos fueron normalizados antes de realizar el proceso.

A diferencia del procedimiento utilizado para la selección de las variables en los vapores simples, en el conjunto de medidas de los vapores simples y sus mezclas binarias el proceso de selección se ha realizado de dos formas diferentes: en la primera, igual que en el análisis anterior se implementaron los diferentes métodos de selección de variables (secuenciales y estocásticos) incluyendo el método estocástico simulated annealing. Por otro lado, el segundo procedimiento se ha realizado en dos pasos, en el primero se aplicó un rápido proceso de selección utilizando el criterio de la varianza para reducir drásticamente el número de las variables. En un segundo paso, que tiene lugar sobre el subconjunto de variables resultantes del primero, se utilizaron procedimientos de selección más finos, basados en los métodos secuenciales y estocásticos mencionados anteriormente.

4.2.6.1 Proceso de selección realizado en una sola fase

Inicialmente, se construyeron los modelos fuzzy ARTMAP y PNN empleando la totalidad de las variables, El número de entradas fue de 120 porque son el número de parámetros extraídos de los 12 sensores. El número de neuronas en la capa de salida fueron 24, correspondiéndose con la presencia de 24 clases de medidas (3 vapores simples y 3 mezclas binarias a cuatro concentraciones diferentes). El porcentaje de aciertos en la identificación y cuantificación se estimó utilizando una validación cruzada de orden uno usando las 96 medidas disponibles. El porcentaje de aciertos en la identificación fue de 90.62% para la fuzzy ARTMAP y 87.5% para la PNN. A partir de estos resultados iniciales, se determinó como influían los procesos de

ISBN:978-84-693-4070-7/DL:T-1167-2010

selección de variables previos al entrenamiento y evaluación mediante este tipo de redes.

Los procesos de selección de variables respectivos se realizaron utilizando la matriz de selección. Los resultados obtenidos de validación son relativamente buenos con unos porcentajes de acierto en la identificación de 90.62% para el forward y backward y un 89.58% para el stepwise acoplados al modelo fuzzy ARTMAP, con un número de variables seleccionadas de 4, 67 y 69 respectivamente. En el caso de la red PNN los resultados fueron de un 91.66, 88.54 y 91.66% para la forward, backward y stepwise selection con un número de parámetros seleccionados de 9, 23 y 25 respectivamente.

Es importante resaltar que los porcentajes de acierto en la identificación y cuantificación de los compuestos volátiles con las variables seleccionadas en los procesos de selección se han estimado usando las 96 medidas (es decir el conjunto inicial incluyendo la matriz de validación) utilizando validación cruzada de orden uno. El hecho de que los clasificadores fueron validados usando medidas diferentes a las empleadas en el proceso de selección de variables provenientes de la matriz de validación prueba la capacidad y la habilidad de generalización de los métodos desarrollados. Cabe recordar que, al igual que en el caso de la identificación de los vapores simples, con estos métodos no se está seguro de sí el subconjunto de variables seleccionadas es el óptimo ya que se puede haber estancado la búsqueda en un mínimo local.

Se utilizó una vez más un GA acoplado a los modelos clasificadores utilizados hasta ahora para seleccionar entre las 120 variables. Este proceso también fue realizado empleando la matriz de selección definida anteriormente. Con un PRE muy parecido, se selecciona el modelo que emplee menos variables. Por consiguiente, se seleccionan 47 de las 120 variables. Un segundo GA se aplicó a las 47 variables seleccionadas en el primer GA. El algoritmo converge después de 86 iteraciones. Después de esta segunda selección, 17 y 14 variables se seleccionaron para cada modelo.

Estas variables se utilizaron para construir los modelos fuzzy ARTMAP y PNN. La tasa de aciertos de clasificación utilizando el conjunto de validación fue de 91.66

% para ambos modelos. Estos resultados se mostraron favorables comparados con los obtenidos sin ningún método de selección y empleando la totalidad de las variables.

Con el método estocástico Simulated Annealing (SA) se acabaron seleccionando 12 y 14 variables, utilizando, como en los casos anteriores, redes fuzzy ARTMAP y PNN como clasificadores. Empleando estas variables en la clasificación se obtuvieron porcentajes de acierto en la clasificación de 89.2 y 90.62% respectivamente. El algoritmo SA se implementó con 50 temperaturas diferentes y el número de iteraciones por temperatura fue de 119. Los resultados de clasificación obtenidos usando los métodos de selección estocásticos fueron similares a los obtenidos sin ningún proceso de selección. Sin embargo, lo más destacable es que el número de variables fue reducido en un factor de 10.

Método SV	# Variables seleccionadas	% de aciertos
Sin método de selección Validación con fuzzy ARTMAP	120	90.62%
GA-fuzzyartmap	17	91.66%
SA-fuzzy artmap	14	89.2%
Forward-fuzzy artmap	4	90.62%
Backward-fuzzy artmap	67	90.62%
Stepwise-fuzzy artmap	69	89.58%
Sin método de selección Validado PNN	120	87.5%
GA-PNN	14	91.66%
SA-PNN	12	90.62 %
Forward-PNN	9	91.66%
Backward-PNN	23	88.54%
Stepwise-PNN	25	91.66%

Tabla 4.4 Resultados de la validación utilizando todos los métodos desarrollados de selección de variables y la totalidad de las variables.

ISBN:978-84-693-4070-7/DL:T-1167-2010

La tabla 4.4 muestra los resultados de las diferentes técnicas de selección de variables estudiadas. El inconveniente más importante encontrado en las técnicas aplicadas ha sido el tiempo de ejecución necesario para cada uno de ellos, especialmente en los métodos SBS, stepwise y GA.

METODOS DE SELECCION	VARIABLES SELECCIONADAS	
	Parámetro	Sensor
GA- Fuzzy Artmap	G_f	825,882,2611
	ΔG	825,822, 800
	ΔG_n	2620
	G_m	825, 882,882,2611
	ΔG_m	825,822,2620,825
	$t_{30–60}$	882-2611
SA-Fuzzy Artmap	G_i	2620
	G_f	825,2620
	ΔG	800,882
	G_m	882,882
	ΔG_m	822,882,822
	$t_{30–60}$	2611
SFS-Fuzzy Artmap	G_f	825
	ΔG	822,800
	G_m	2611
SFS-PNN	G_i	800,882
	G_f	800,2611
	ΔG	826,882
	G_m	2620
	ΔG_m	882
	t_m	882

Tabla 4.5.a Lista de variables seleccionadas con los métodos implementados utilizando la totalidad de las variables (sin el criterio de la varianza)

METODOS DE SELECCION	VARIABLES SELECCIONADAS	
	Parámetro	Sensor
GA-PNN	G_i	2620
	G_f	882,2611
	ΔG	825,882
	G_m	882,2160,2611,2620
	ΔG_m	822
	$\Delta G_{m,n}$	2620
	$t_{30–60}$	822,822,882
SA-PNN	G_i	826
	G_f	825,2620
	ΔG	2620
	G_m	2611
	ΔG_m	822
	$\Delta G_{m,n}$	822,882
	$t_{30–60}$	822,2611
	$t_{10–90}$	2620

Tabla 4.5.b (continuación) Lista de variables seleccionadas con los métodos implementados utilizando la totalidad de las variables (sin el criterio de la varianza)

La tabla 4.5 muestra que parámetros fueron seleccionados por los métodos empleados. La mayoría de los métodos han seleccionado parámetros similares, especialmente G_i, G_f, ΔG, G_m, ΔG_m y $t_{30–60}$. Sin embargo, muchos de los sensores seleccionados difieren dependiendo del método de selección empleado. Los sensores utilizados son conocidos por su alta colinealidad por lo que las 120 variables disponibles presentan mucha colinealidad entre ellas, especialmente aquellas de diferentes sensores pero de misma naturaleza. Consecuentemente, el problema de optimización considerado aquí no tiene una solución única. Es decir, que utilizando diversos subconjuntos de variables nos conducirá a resultados muy similares en la

ISBN:978-84-693-4070-7/DL:T-1167-2010

identificación y cuantificación de los gases. Esto explica porque con diferentes métodos (secuenciales y/o estocásticos) se escogen diferentes subconjuntos de variables que proporcionan resultados similares en la clasificación.

4.2.6.2 Proceso de selección empleando dos fases.

Un gran problema presentado en la realización de las pruebas anteriores fue el tiempo de cálculo empleado para la selección utilizando los diferentes métodos, siendo éste un claro inconveniente ya que el constante incremento tanto de medidas como de parámetros produce un mayor consumo en el tiempo de proceso. Por este motivo se implementó un nuevo procedimiento para reducir el tiempo de cálculo en un factor 4 aproximadamente.

En el nuevo procedimiento se distinguen dos partes: primero, se aplica una técnica de selección rápida aplicando un criterio de mérito a cada variable (FM). Esta técnica basa su cálculo en la relación entre la varianza intraclase y la varianza interclase de las variables. Una variable con baja varianza dentro de una clase (un tipo de compuesto volátil) y alta varianza entre clases (diferentes compuestos) implica que el parámetro en cuestión presenta buena repetitividad y alta selectividad, lo cual es muy interesante para poder discriminar correctamente entre las clases preestablecidas. Esta técnica de selección se explica más detalladamente en el capítulo 3, específicamente en el subapartado (3.5.1).

En este método se fija un valor umbral (límite) para el factor de mérito y solamente las variables con un factor de mérito superior al umbral se conservan para la segunda fase de selección. En la segunda fase es cuando se aplican los métodos secuenciales (forward selection, backward elimination y stepwise selection) o estocásticos (Algoritmos Genéticos y Simulated Annealing) comentados anteriormente.

Al realizar el primer paso con la técnica de determinación del factor de mérito por medio del criterio de la varianza usando un valor umbral de 3 (este valor se determinó analizando el resultado del criterio de la varianza, de forma que las

variables seleccionadas fueran las menores posibles, pero sin llegar a una reducción exagerada), se obtuvo una reducción de las 120 variables iniciales a 31 con un tiempo de cálculo muy reducido (es decir cerca del 75 % de las variables originales se descartaron). Tomando como base este número de variables se pasó a realizar la selección con cada uno de los métodos anteriormente mencionados acoplándolos a los modelos clasificadores.

(FM) solo validado fuzzy ARTMAP	31	88.54%
GA-fuzzyartmap	12	91.66%
SA- fuzzyartmap	11	89.2%
Forward-Fuzzy artmap	7	90.62%
Bacward-Fuzzy artmap	11	90.62%
Stepwise-fuzzy artmap	13	90.62%
(FM) solo Validado PNN	31	83.33%
GA-PNN	11	92.70%
SA-PNN	9	91.66%
Forward-PNN	6	88.54%
Backward-PNN	8	90.62%
Stepwise-PNN	10	92.70%

Tabla 4.6 la selección de variables realizada empleando los 2 pasos. El criterio de la varianza y el proceso de selección fino con los diferentes métodos de selección

Los resultados que se obtienen son muy parecidos a los obtenidos anteriormente sin emplear este procedimiento (ver tabla 4.6) con la ventaja de que el tiempo empleado para realizar estas selecciones se reduce en un factor 4 aproximadamente, comparado con el tiempo de cálculo empleado con todas las variables en una sola etapa.

METODOS DE SELECCION	VARIABLES SELECCIONADAS	
	Parámetro	Sensor
GA- Fuzzy Artmap	G_f	800
	ΔG	822,800, 882
	G_m	825,800,882,2620
	ΔG_m	825, 822,882,2620
SA-Fuzzy Artmap	G_f	825,825,2620
	ΔG	825,822,822
	G_m	826,882,2620
	ΔG_m	822,882
	$t_{30–60}$	800
SFS-Fuzzy Artmap	G_f	825,800
	ΔG	800,822
	G_m	882,825,825
SFS-PNN	G_f	825
	ΔG	825,822,822,2620
	G_m	882
GA-PNN	G_f	2620
	ΔG	822,822,882,2620
	G_m	822,800,882,2620
	ΔG_m	822,822
SA-PNN	G_f	825,800,2620
	ΔG	882
	G_m	2620
	ΔG_m	825,825,822

Tabla 4.7 Lista de variables seleccionadas con los métodos implementados utilizando previamente el criterio de la varianza como proceso de preselección).

La tabla 4.7 muestra qué variables fueron seleccionadas con los métodos empleados en esta nueva estrategia. Muchos de los métodos seleccionan variables muy similares, especialmente G_f, ΔG, G_m y ΔG_m. En este nuevo proceso hay también

una coincidencia notable entre los sensores seleccionados por los diversos métodos. Esto puede ser debido a la reducción en la dimensionalidad (de 120 a 31) en el primer paso de la selección (intra/intervarianza).

Finalmente, como conclusión de esta sección podemos decir que acoplando redes clasificadoras (fuzzy ARTMAP y PNN) a diferentes métodos de selección secuenciales y estocásticos, proporcionan unos buenos resultados en los procesos de selección de variables en problemas de análisis de gases en sistemas multisensoriales como el utilizado en nuestro caso.

Los algoritmos de selección tales como el forward, backward y el stepwise pueden ser usados con éxito, pero cuando hay una gran cantidad de posibles soluciones pueden quedar atrapados en mínimos locales en el proceso de optimización. Por este motivo se optó por probar y comparar dicha metodología con algoritmos estocásticos como los algoritmos genéticos y el simulated annealing. Aunque los GA y el simulated annealing pueden encontrar buenas soluciones en las aplicaciones consideradas (casi siempre mejores que con métodos secuenciales) presentan el inconveniente de la gran carga computacional que requieren. Una posible solución a este dilema consistió en introducir un paso previo de preseleccion por medio del criterio de la varianza para reducir drásticamente el número de variables en un tiempo considerable. Los resultados obtenidos y descritos en esta sección han sido publicados en dos artículos [3,4].

4.3 Selección de variables para aplicaciones de sistemas de olfato electrónico basados en espectrometría de masas.

4.3.1 Introducción

En este subapartado se estudian de nuevo diferentes métodos de selección de variables aunque en este caso para ser aplicados a narices electrónicas basadas en espectrometría de masas (MS). Debido a las nuevas propiedades de los datos

ISBN:978-84-693-4070-7/DL:T-1167-2010

obtenidos con este tipo de instrumentos, en esta parte del estudio se introducen nuevas estrategias en lo que a selección de variables se refiere [5].

Para evaluar estas nuevas estrategias, se generaron tres conjuntos de medida obtenidos con un espectrómetro de masas configurado para actuar como un sistema de olfato electrónico. En esta configuración, la columna cromatográfica del instrumento GC-MS original es desactivada manteniendo una temperatura de 250º durante todo el proceso de análisis. En algunas referencias se denomina a esta técnica como "Espectrometría de Masas Directa".

La selección de variables fue realizada en tres pasos. En los dos primeros pasos el objetivo principal es eliminar variables con poca información y alta colinealidad (i.e, redundancia) de forma secuencial. Estos dos pasos son computacionalmente rápidos y permiten reducir drásticamente el número de variables descriptoras en las medidas (cerca del 80 % de las variables iniciales son eliminadas al finalizar el segundo paso). En el tercer paso se utilizó alguno de los métodos de selección de variables tratados en esta tesis como, por ejemplo, el método simulated annealing (SA) el cual continua reduciendo significativamente el número de variables sin reducir la tasa de éxito de los algoritmos de reconocimiento de patrones.

4.3.2 Conjuntos Experimentales

4.3.2.1 Conjunto de muestras de solventes

La primera base de datos consistió en medidas de mezclas de disolventes orgánicos utilizando un headspace autosampler. El conjunto de medidas estaba compuesto por 4 soluciones diferentes de etanol puro con impurezas agregadas (trichloroetileno, 1-butanol, etilbenceno y tolueno). La composición exacta de las diferentes muestras se puede ver en la tabla 4.8. Para las cuatro soluciones iniciales, se obtuvieron 6 diferentes muestras (seis soluciones fueron preparadas en 6 diferentes frascos). Para cada una de estas 6 muestras se tomaron y almacenaron 5 diferentes alícuotas de 10 ml dentro de viales de 20 ml y se sellaron herméticamente con septa de silicona. (Es decir en total se analizaron 120 viales.)

	Trichloroetileno	1-Butanol	Etilbenceno	Tolueno
S1	1	1	0	1
S2	1	1	1	1
S3	1	1	1	0
S4	1	0	1	1

Tabla 4.8 Composición de las diferentes muestras en el conjunto 1. Las cantidades son expresadas en % disuelto en etanol.

El muestreo basado en micro extracción en fase sólida (SPME) fue realizado con una fibra Carboxen/PDMS de 75- μm, comprada a Supelco (Supelco Park, Bellefonte, PA). Antes de cualquier extracción, la fibra se acondicionó siguiendo las recomendaciones del fabricante. En cada medida, la fibra fue desprotegida de su cubierta de acero inoxidable y la muestra expuesta al espacio de cabeza de los diferentes viales durante 20 minutos a temperatura ambiente. El SPME se ajustó a 3.0 unidades de la escala para asegurarse de que la fibra fuese colocada sobre la muestra en el headspace de la misma, pero sin llegar a tener contacto físico con la muestra líquida.

Para implementar la nariz electrónica basada en MS se usó un equipo Shimadzu QP 5000 GC/MS (Shimadzu Corp., Tokyo, Japan). En el instrumento se instaló una columna capilar (Supelcowax, 30m × 0.25 mm i.d., × 0.25 mm de grosor de capa). Los compuestos volátiles atrapados en la fibra SPME fueron posteriormente desorbidos durante 3 minutos a 280° C dentro del inyector del cromatógrafo de gases (CG). El gas portador que se empleó fue helio puro en un 99.995%. La temperatura del CG y de la interfaz CG/MS se mantuvo a 250°C para evitar la separación

ISBN:978-84-693-4070-7/DL:T-1167-2010

cromatográfica (obteniendo, de esta forma, un pico coeluido). Para cualquier medida dada cada pico es integrado y el espectro de masas resultante proporciona una huella digital que es característica del headspace de la muestra analizada.

Los espectros de masa fueron registrados a una razón de 2 scans/s sobre los cocientes de m/z entre 40 y 150 amu. Este rango es conocido por contener todos los fragmentos m/z con las intensidades más altas para los compuestos implicados en esta base de datos. La totalidad de los iones obtenidos bajo estas condiciones toman la forma de un pico asimétrico como se muestra en la figura 4.3. La matriz respuesta contiene los valores de los fragmentos totales registrados entre el scan 1 y el scan 70 para cada muestra analizada. El análisis de datos se realizó sobre los espectros de masa relativos (es decir, normalizado por la amplitud del pico más alto).

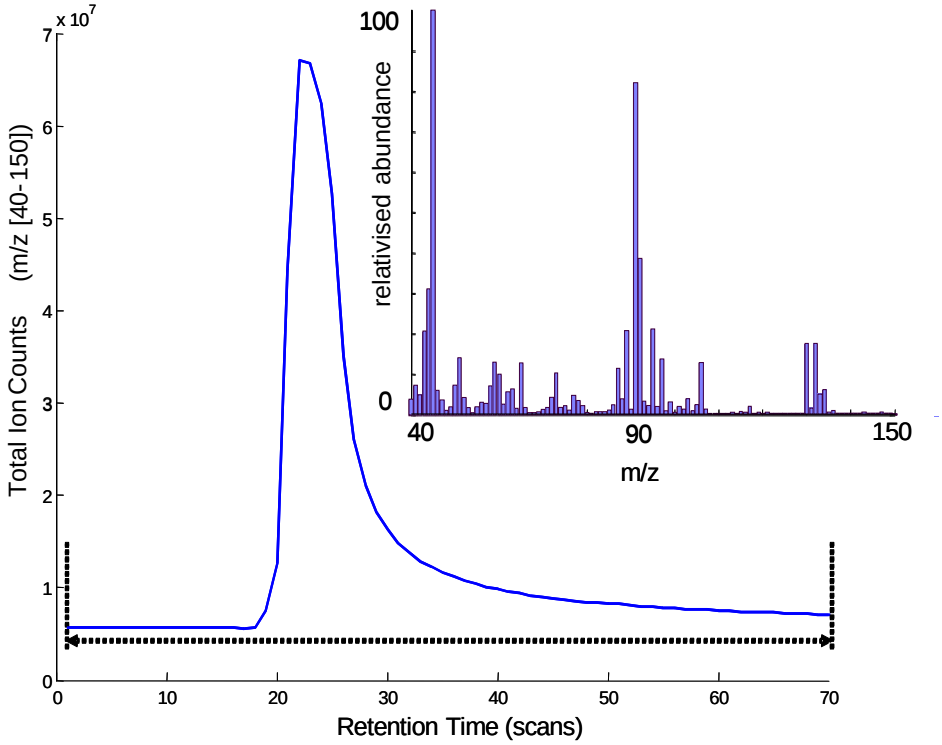

Figura 4.3 Pico cromatográfico coeluido de la muestra S2 obtenida por un SPME-MS.

4.3.2.2 Análisis del conjunto de solventes

A priori, el principal desafío para poder identificar correctamente estos compuestos se debe a la presencia del etilbenceno y del tolueno en las mezclas, pues estas dos especies muestran algunas semejanzas entre sus espectros de masas. La tabla 4.9 muestra cuáles son los fragmentos m/z más intensos encontrados en el espectro de masas de todos los compuestos usados. Este conjunto puede considerarse como una prueba de referencia para evaluar los diferentes métodos de selección de variables.

El objetivo principal que se buscó con este conjunto fue mirar si con los tres pasos empleados para la selección de las variables podría determinarse correctamente qué relaciones masa- carga serían las correctas para identificar las mezclas. Antes de realizar la selección de variables, una red fuzzy ARTMAP se entrenó y validó usando validación cruzada de orden uno. La clasificación utiliza 111 entradas y el número de categorías fue de 4 que son el número de las diferentes mezclas analizadas. El porcentaje de aciertos en la clasificación fue de 95.83 % que corresponde a un solo error, concretamente una muestra de S2 que fue identificada incorrectamente como S4. (Ver tabla 4.9).

Compuesto	10 fragmentos m/z mas intensos
Tricloroetileno	132, 130, 95, 97, 60, 134, 47, 62, 59, 94
1-Butanol	56, 41, 43, 42, 55, 45, 40, 57, 44, 53
Etilbenceno	91, 106, 51, 65, 77, 78, 50, 92, 52, 63
Tolueno	91, 92, 65, 51, 63, 45, 50, 46, 62, 89

Tabla 4.9 Los fragmentos m/Z mas intensos encontrados en el espectro de masa de los diferentes compuestos usados en el conjunto de datos de los solventes. Los fragmentos aparecen clasificados por orden de decrecimiento de intensidad.

El conjunto formado por las 120 medidas fue dividido en dos matrices de 100 y 20 filas respectivamente. La primera matriz contiene 25 medidas replicadas por

ISBN:978-84-693-4070-7/DL:T-1167-2010

mezcla solvente, mientras que la segunda contiene las 5 medidas restantes por solvente. El primer subconjunto (matriz de selección) se utilizó para realizar el proceso de selección de las variables y el segundo subconjunto (matriz de validación) fue empleado para validar los resultados. Este proceso se repite 6 veces variando los subconjuntos de selección y validación.

El proceso de selección de variables se realizó en tres pasos. En los dos primeros pasos el objetivo principal fue el de eliminar las variables (relaciones m/z) con poca información y alta colinealidad (i.e, redundancia) de forma secuencial. Inicialmente, eliminamos aquellas variables con información ruidosa utilizando la técnica del criterio de la varianza (explicado en el capítulo 3), fijando un umbral que nos permita poder discriminar entre las variables. Dicho valor fue de 0.5, valor que permitió seleccionar un número reducido de variables con un valor de mérito superior a este umbral, seleccionando entre 29 y 37 de las 111 variables iniciales

El segundo paso empleado fue el de eliminar variables colineales fijando en 0.15 el valor umbral, este valor permite seleccionar un número reducido de variables con valores de colinealidad inferiores a este umbral (esta técnica es explicada con mayor profundidad en el capítulo 3). Con este paso, 20 variables quedaron sin eliminar. Es importante destacar que el tiempo de calculo de los dos primeros pasos empleado fue de aproximadamente 6 minutos en una plataforma PC basada en Pentium 4.

Una vez hecha la pre-selección, se ejecutó el método de selección estocástico simulated annealing para seleccionar las variables más importantes entre las restantes de los dos primeros pasos. Los algoritmos SA se ejecutaron empleando 50 diferentes temperaturas de recocido y el número de iteraciones por temperatura fue de 17 (mas detalles sobre este algoritmo pueden encontrarse en el capítulo tres de esta tesis). Al final del proceso solo 3 de las variables se mantuvieron seleccionadas, concretamente las relaciones m/z 46, 56 y 106. El porcentaje de acierto en la clasificación de la mezcla de los solventes estimada sobre los 6 conjuntos entrenamiento/validación fue del 100 %. La figura 4.4 muestra un diagrama de bloques del proceso de selección y de validación. Es importante tener presente que para cada clasificador fuzzy ARTMAP, la validación implica usar medidas que no han participado en el proceso de selección de variables y son, por lo tanto, medidas nuevas.

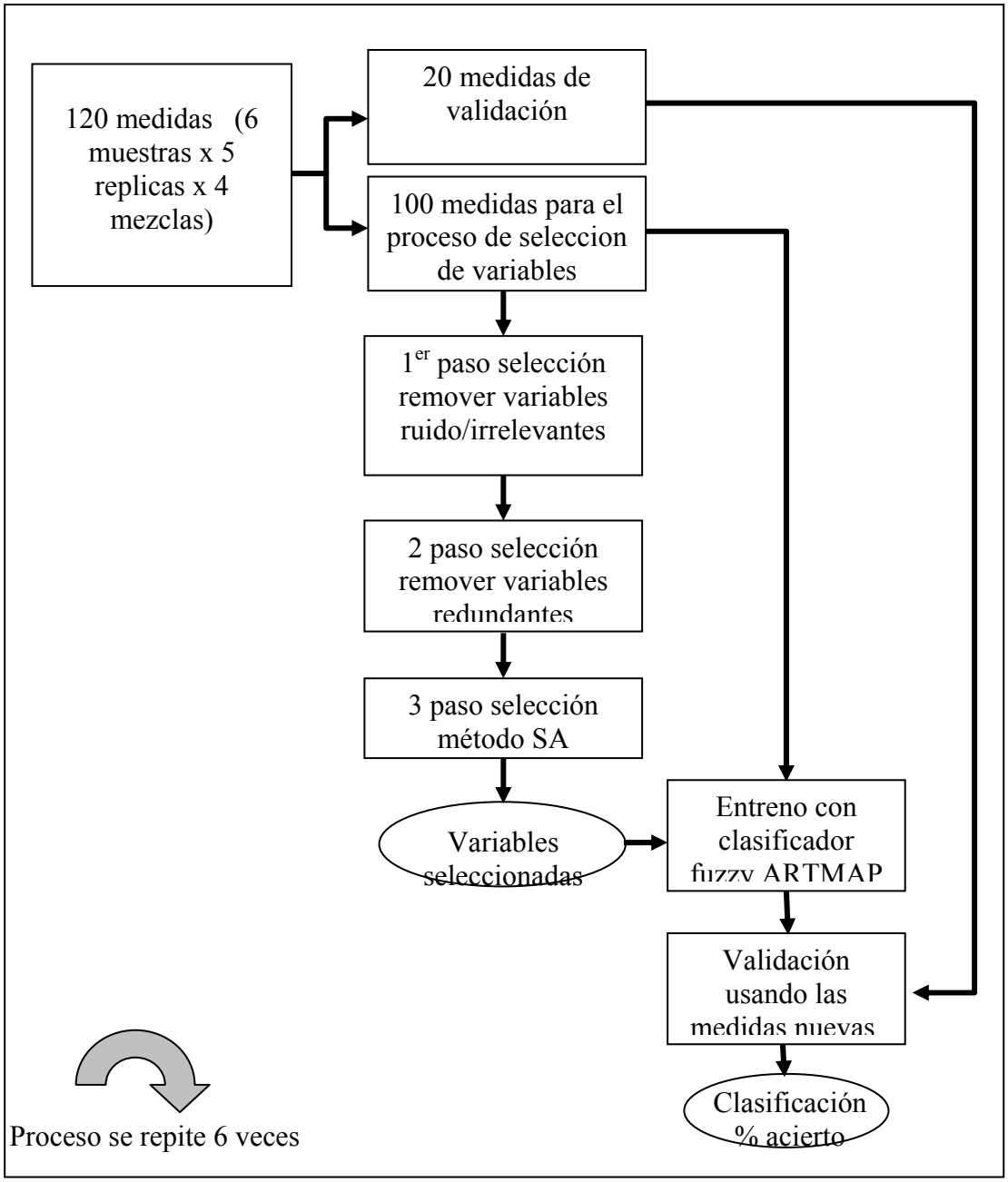

Figura 4.4 Diagrama de bloques de los procesos de selección de las variables y validación para cada uno de los bloques.

ISBN:978-84-693-4070-7/DL:T-1167-2010

Considerando la tabla 4.11 puede deducirse que las variables seleccionadas tienen las siguientes especificaciones:

- m/z=46, es la masa más relevante para el tolueno, ya que no se encuentra en los otros compuestos de la mezcla solvente.
- m/z=56, es la masa más relevante para el 1-Butanol.
- m/z=106, la segunda masa más relevante para el Etilbenceno.

Es importante decir que los diferentes procesos de selección de variables presentan cierta indecisión respecto a la m/z = 91, ya que aún siendo el ion más frecuente para el etilbenceno y el tolueno a la vez, y por lo tanto no sirve para ayudar a discriminar entre estos dos compuestos. El último comentario a resaltar es que no se ha seleccionado ninguna relación masa-carga característica del tricloroetileno. Esto es correcto porque el tricloroetileno está presente en todas las muestras que se desean discriminar, y, por lo tanto, su caracterización no es necesaria para una buena discriminación entre las muestras analizadas. (Véase tabla 4.9). Estos resultados muestran que el proceso con los diferentes pasos de selección de variables introducidos aquí es apto para hallar la información esencial necesaria para resolver el problema de discriminación considerado. El tiempo total empleado en el proceso completo de la selección requirió alrededor de 15 minutos en una plataforma PC Pentium 4.

4.3.2.3 Conjunto de medidas de aceites de oliva:

Para este conjunto se recopilaron nueve muestras de aceite de oliva virgen de la región de Tarragona (España). Las muestras del aceite se recibieron en botellas de vidrio transparente de 500-ml y se conservaron en un refrigerador a 4ºC. El día antes de que se analizaran, las muestras se sacaban del refrigerador y se mantenían en oscuridad a temperatura ambiente para que se fuesen calentando a la temperatura ambiente.

Las primeras 3 muestras del aceite (variedad de oliva empeltre) se clasificaron como aceite de oliva extra virgen por un panel sensorial. Las olivas fueron molidas

en el mismo molino y el aceite almacenado en tres tanques diferentes. Las tres muestras siguientes (variedad de oliva 'arbequina') también fueron clasificadas como aceite de oliva extra virgen por el panel sensorial. Las olivas se trituraron en tres molinos diferentes y fueron almacenadas en diferentes tanques. Finalmente, las últimas tres muestras (variedad de oliva 'arbequina'') se trituraron en el mismo molino y fueron almacenadas en diversos tanques. Estas ultimas fueron clasificadas como defectuosas por el panel sensorial, el cual considera que la muestra 7 y 8 mostraron defectos leves y la muestra 9 presentaba defectos graves (aceite lampante) (véase la tabla 4.10).

	Variedad	Molino	Tanque almacenamiento	Evaluación sensorial
S1	EM	A	1	EV
S2	EM	A	2	EV
S3	EM	A	3	EV
S4	AR	B	4	EV
S5	AR	C	5	EV
S6	AR	D	6	EV
S7	AR	E	7	SD
S8	AR	E	8	SD
S9	AR	E	9	LA

Tabla 4.10. Estudio de las muestras de aceite de oliva. EM: empeltre, AR: arbequina, EV: extra virgen, SD: levemente defectuosa, LA: lampante

Cinco partes alícuotas de 10 ml de cada muestra del aceite de oliva se extrajeron con una pipeta de las botellas de cristal de 500 ml, y fueron colocadas en viales de cristal de 20 ml sellándolas herméticamente con septum de silicona (es decir 45 viales en total). Se trató de evitar la agitación de las botellas de vidrio durante el

ISBN:978-84-693-4070-7/DL:T-1167-2010

proceso de pipeteo debido a que las muestras transferidas dentro de los viales tienen que ser representativas de los aceites guardados en los tanques.

Cuatro de los cinco viales de muestras de aceite de oliva (es decir, 36 frascos) se utilizaron para realizar las medidas usando el sistema de olfato electrónico basado en MS. De nuevo, la técnica SPME fue utilizada para el muestreo. Las condiciones usadas para la entrega de muestras, el registro del espectro de masas y la normalización de las mismas, fueron idénticas a las descritas en la base de datos anterior (conjunto de solventes). La única diferencia estuvo en el rango del espectro (en este caso entre 45 y 150 amu) de las relaciones m/z que se exploraron para esta aplicación. En este rango se concentra la mayor información espectral relacionada con las medidas de aceites de oliva. Los fragmentos del 40 al 44 no se tuvieron en cuenta y fueron ignorados debido a que en estos fragmentos es más significativa la presencia del gas portador que de la muestra de los aceites en sí.

El quinto vial de cada muestra de aceite de oliva se utilizó para evaluar los perfiles cromatográficos. Dichos perfiles se obtuvieron usando el equipamiento descrito anteriormente. Se utilizó una secuencia de variación de la temperatura para alcanzar la separación cromatográfica de los compuestos volátiles presentes en el headspace de las muestras de aceite de oliva.

Inicialmente, el CG se mantuvo a una temperatura de 40ºC durante 3 minutos para luego ser elevada hasta 70ºC a un ritmo de 2ºC/min. Finalmente, dicho ritmo se aumentó alcanzando los 10ºC/min hasta llegar a una temperatura de 250ºC. El detector de masas estuvo operando en el modo de impacto de electrones (70 eV) con un rango espectral de 45 a 150 amu. La temperatura de la fuente de iones se mantuvo en 250ºC.

4.3.2.4 Análisis del conjunto de aceites

Esta base de datos es más complicada que la anterior ya que pertenece a muestras reales que producen espectros de masa mucho más complejos. Además, el número de categorías posible es más alto, concretamente nueve.

La figura 4.5 muestra el cromatograma que se obtuvo para las 9 muestras de aceite de oliva estudiadas. El cromatograma aparece agrupado en la figura de acuerdo a la variedad del aceite. Los componentes fueron identificados mediante su espectro de masas usando el NIST y las librerías Wiley. El resultado de este proceso de identificación se muestra resumido en la tabla 4.11. Esta tabla muestra que los perfiles de los volátiles son bastantes más complejos y diferentes que los del conjunto sintético. Por consiguiente, cada fragmento m/z medido puede ser originado por la ionización de varios compuestos diferentes presentes en el headspace del aceite.

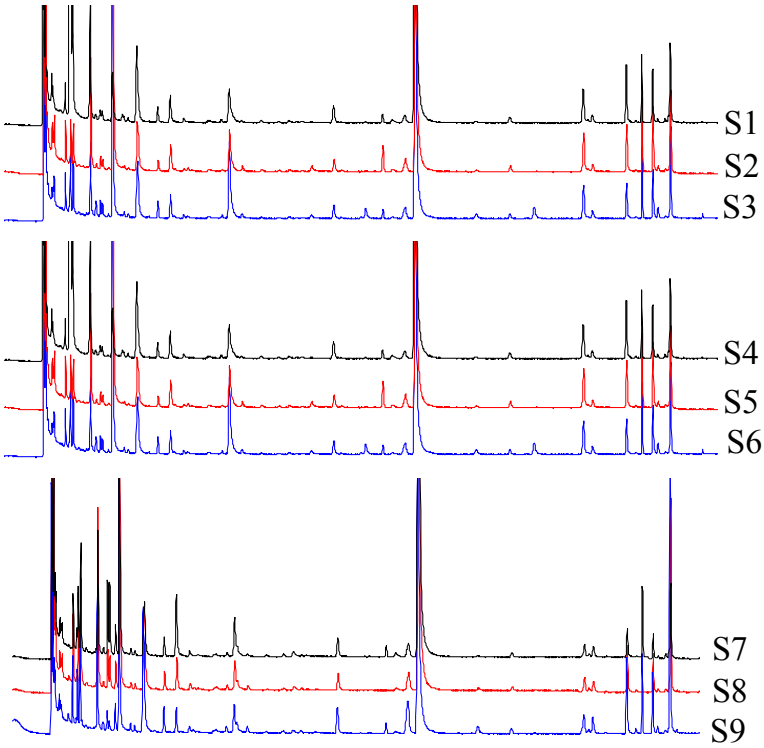

Figura 4.5 perfiles cromatográficos para las 9 diferentes muestras de aceite de oliva. Obtenidos de la señal proveniente del CG-SM

ISBN:978-84-693-4070-7/DL:T-1167-2010

	S1	S2	S3	S4	S5	S6	S7	S8	S9
Hydrocarbonos									
Pentane	9.1	9.1	6.8	1.8	5.5	7.7	4.8	5.6	3.9
2-methyl-1-Butene	ND	ND	ND	1.5	2.5	ND	ND	ND	ND
2,3-dimethyl- butane	ND	10.4	7.9	ND	ND	ND	ND	ND	ND
2-methyl- Pentane	21.4	ND	ND	ND	ND	ND	ND	ND	ND
Hexane	ND	10.4	7.3	ND	ND	ND	ND	ND	ND
1,4-Pentadiene	ND	ND	ND	1.0	1.7	1.5	0.8	1.1	0.7
2,3-dimethyl-Pentane	2.02	ND	1.55	ND	ND	ND	ND	ND	ND
Heptane	ND	ND	1.2	ND	ND	ND	ND	ND	ND
Octane	ND	ND	ND	0.3	0.9		1.3	1.6	1.4
Toluene	1.9	1.5	1.8	ND	ND	ND	ND	ND	ND
Total	**32.4**	**31.4**	**25.0**	**4.6**	**10.6**	**9.2**	**7.0**	**8.4**	**6.0**
Alcoholes									
3-Methyl-1-butanol	ND	ND	ND	ND	ND	ND	ND	5.10	ND
Isopropyl alcohol	7.4	7.2	5.6	ND	ND	ND	1.0	1.0	0.9
Ethanol	1.1	1.4	2.1	1.3	36.8	23.5	8.6	12.4	11.0
3-Hepten-1-ol	ND	ND	ND	0.7	ND	1.2	ND	1.5	ND
1-Hexyn-3-ol	ND	ND	ND	ND	1.3	ND	ND	ND	ND
1-penten-3-ol	ND	ND	ND	0.5	0.6	0.7	0.9	0.99	1.4
2-methyl-1-butanol	ND	ND	ND	ND	ND	ND	ND	ND	2.6
3-Methyl-1-butanol	ND	ND	ND	ND	ND	ND	ND	1.4	ND
1-hexanol	0.9	0.9	1.3	1.3	1.7	1.5	1.1	1.9	2.8
3-hexen-1-ol	1.1	1.0	1.3	1.1	2.2	2.1	0.8	1.4	2.0
2-hexen-1-ol	0.4	ND	0.7	1.4	2.3	3.3	2.1	6.1	12.6
Total	**10.9**	**10.5**	**10.9**	**6.3**	**44.9**	**32.3**	**14.5**	**25.6**	**33.2**
Aldehydos									
Acetaldehyde	1.3	1.3	ND	ND	1.1	1.0	1.3	0.7	0.3
Propanal	2.2	ND	ND	ND	ND	ND	ND	ND	ND

Tabla 4.11.a Compuestos volátiles encontrados en el headspace del aceite de oliva para las 9 muestras (S1-S9) (expresado como unidades de porcentaje de área)

	S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**
Aldehydos									
2,4-dimethyl-Pentanal	ND	ND	ND	ND	1.0	ND	ND	ND	ND
2,3-dimethyl-Pentanal	ND	2.0	ND	ND	ND	ND	ND	ND	ND
2-methylhexanal	ND	ND	1.4	ND	ND	ND	ND	ND	ND
2-Methylbutanal	ND	ND	ND	ND	0.4	0.6	2.2	1.3	0.5
3-Methylbutanal	ND	ND	ND	ND	0.3	0.5	2.3	1.1	0.2
Pentanal	0.5	ND	1.4	ND	ND	4.5	ND	ND	ND
2-hexenal	8.6	11.6	21.8	19.1	20.4	23.6	41.8	30.5	24.4
nonanal	ND	ND	ND	ND	ND	0.4	ND	ND	0.3
Total	**18.0**	**20.1**	**31.9**	**20.7**	**25.1**	**36.9**	**50.0**	**35.5**	**28.0**
Ketones									
Acetone	10.6	9.1	5.7	53.1	1.0	1.7	1.5	2.3	1.3
3-Pentanone	ND	ND	ND	2.7	3.0	ND	ND	5.6	7.2
1-penten-3-one	ND	ND	ND	ND	ND	ND	3.0	ND	ND
Total	**10.6**	**9.1**	**5.7**	**55.8**	**4.0**	**1.7**	**4.5**	**7.9**	**8.5**
Esters									
Methyl acetate	ND	ND	ND	2.1	0.9	1.3	2.5	2.4	2.1
Ethyl acetate	0.7	0.6	0.5	2.5	3.1	2.6	3.4	5.4	3.8
3-hexen-1-ol-acetate	1.6	1.9	2.1	1.0	1.7	1.8	0.4	0.5	1.0
Total	**2.3**	**2.5**	**2.6**	**5.5**	**5.7**	**5.8**	**6.3**	**8.3**	**6.9**

Tabla 4.11.b (Continuación) Compuestos volátiles encontrados en el headspace del aceite de oliva para las 9 muestras (S1-S9).

Inicialmente se realizó una clasificación con las 9 categorías empleando la red fuzzy ARTMAP sin los pasos previos de selección de variables. Como entradas para la clasificación se utilizó el rango entre 45 y 150 de relaciones m/z, La capacidad de acierto en la clasificación de los aceites de oliva se estimó en un 94.44% usando una validación cruzada de orden uno.

Sólo existieron confusiones entre 2 muestras de aceite de oliva: una muestra S1 que fue clasificada como S3 y una muestra S3 que se clasificó como S1. Estas muestras de aceites derivan del mismo tipo de aceite (‘empeltre’), aunque las olivas

ISBN:978-84-693-4070-7/DL:T-1167-2010

habían sido trituradas en el mismo molino en diferentes días y almacenadas en diferentes tanques.

En un proceso similar al empleado en la base de datos de los solventes, el primer paso en la selección de variables fue eliminar ruido o relaciones m/z irrelevantes. El valor umbral que se fijo fue de 0.5 para el criterio intra- varianza/ inter-varianza con lo que el número de variables original (106) se redujo a 48. De igual manera se aplicó el segundo paso (método de colinealidad) para determinar la presencia de variables redundantes y por consiguiente eliminarlas. Se fijo un umbral de 0.7 y el número de variables seleccionadas se redujo a un total de 22. Estas 22 variables fueron utilizadas como variables de entrada para validar y entrenar con una red fuzzy ARTMAP. La tasa de acierto en la clasificación de las muestras (empleando validación cruzada de orden uno) fue otra vez de 94.44%. Los errores se produjeron entre las mismas muestras que en el caso inicial.

En un tercer paso se empleó el método estocástico Simulated Annealing para seleccionar la combinación óptima de parámetros entre las 22 variables que se mantuvieron en los pasos previos. El algoritmo SA se ejecutó con 50 temperaturas diferentes y el número de iteraciones por temperatura fue de 21. Como resultado, sólo 4 de las 32 variables fueron seleccionadas, que corresponden a las relaciones m/z 46, 47, 57 y 58. Nuevamente se utilizo un clasificador fuzzy ARTMAP para entrenar y validar el mismo conjunto de datos inicial empleado para la clasificación de las 9 categorías y el proceso de selección pero solo usando las 4 variables mencionadas anteriormente con un porcentaje de acierto en la clasificación de 100 % (validación cruzada). El proceso completo de selección de las variables requirió un tiempo total de 20 minutos aproximadamente en una plataforma PC Pentium 4.

Aunque el sistema de olfato electrónico basa su funcionamiento en la inspección y el tratamiento matemático del perfil entero en lugar de la asignación de señales individuales a un componente específico, algunos de los fragmentos m/z seleccionados se relacionan a compuestos que ayudan a diferenciar entre las muestras. Así, por ejemplo, la m/z 46 corresponde al pico molecular del etanol (un componente presente en todas las muestras pero en diferentes cantidades) y las m/z 57-58 se corresponden a fragmentos típicos presentes como picos bases o picos con

alta intensidad en muchos de los compuestos volátiles identificados mediante el análisis cromatográfico de los aceites estudiados.

Con estos resultados se puede ver como los métodos de selección de variables escogen parámetros que tienen sentido y pueden ser interpretados químicamente en muchos casos.

4.3.2.5 Conjunto de muestras de jamón ibérico

En este tercer conjunto se analizaron once tipos de jamones ibéricos españoles semicurados. Las muestras se obtuvieron directamente de cinco productores y se diferenciaron en el tipo de alimento que se les suministraba a los cerdos durante su período de cebadura (es decir, bellota o forraje) y en su calidad (tipo de cerdos). La tabla 4.12 da más detalles de los jamones usados.

Marca de jamón (Fabricante)	Nombre corto	# tipo de jamón	Cerdo alimentado con
Extremadura	EX	4	bellota
Guijuelo #1	G1	1	bellota
Huelva	HU	1	bellota
Guijuelo #2	G2	2	forraje
Guijuelo #3	G3	3	forraje

Tabla 4.12 11 tipos de jamones ibéricos españoles analizados la diferencia entre los jamones son el productor el tipo de cerdo y la alimentación dada a ellos.

Las muestras se prepararon de la siguiente forma: tres gramos del jamón (tomado del bíceps femoral) fueron triturados e introducidos en viales de vidrio de 10 ml y sellados con un septum. Diez partes alícuotas se prepararon para cada tipo de jamón (excepto para el jamón producido en Extremadura con tan solo nueve). Por lo tanto, el conjunto comprendía un total de 109 muestras para ser analizadas.

ISBN:978-84-693-4070-7/DL:T-1167-2010

El muestreo se realizó mediante un headspace estático (headspace autosampler Agilent 7694). Las temperaturas de operación del horno, del loop y de la línea de transferencia fueron de 90,100 y 110 ºC, respectivamente. Los tiempos de equilibrado del vial, presurización e inyección fueron de 30, 0.4 y 1 minutos, respectivamente. Las muestras obtenidas en el headspace se introdujeron dentro del inyector del cromatógrafo de gases Hewlett-Packard 6890 series II acoplado a un detector de masas selectivo (Hewlett-Packard HP 5973; Wilmington, DE, USA). El inyector se usó en modo split y se mantuvo a una temperatura de 280ºC. El sistema se equipó con una columna HP 19091J-215 (50m × 0.32mm, con un grosor de capa de 1.05 μm) que se mantuvo a 200ºC en condiciones isotérmicas. De esta forma, la separación cromatográfica se pudo evitar y la columna actuaba únicamente como línea de transferencia para la entrega de volátiles al detector de masas. El flujo en la columna fue de 1.5 ml/min.

Los compuestos volátiles fueron coeluidos dentro del espectrómetro de masas, aparato basado en un detector de impactos de masas electrónico selectivo a 70 eV, con un voltaje de elevación de 2706 V, y con un promedio de recolección de datos de 1 scan s^{-1} con un rango del espectro de 45-250 uma, en este rango se concentra las mayor información espectral relacionada con las medidas de los jamones. Los fragmentos que están por debajo del fragmento 45 fueron ignorados debido a que en éstos fragmentos es más significativa la presencia del gas portador y de aire que de la muestra de jamón en sí.

4.3.2.6 Análisis del conjunto de datos de los jamones ibéricos

Inicialmente se clasificaron las muestras entre 11 categorías posibles utilizando una red fuzzy ARTMAP sin pasos previos de selección de variables. Debido a que en este conjunto el número de medidas disponibles era muy alto, un método diferente de validación se empleó segmentando el conjunto en 5 bloques de validación. Cada conjunto de entrenamiento comprende 8 replicas de las medidas (del total de las 10 disponibles por tipo de jamón) para las 11 categorías, es decir un total de 87 medidas. Hay que aclarar que fueron 87 medidas, ya que para un tipo de jamón se tenían 9

medidas en vez de 10 como en el resto. Los conjuntos de validación correspondientes comprendían 2 medidas por muestra de jamón, diferentes de las del conjunto empleado en la fase de entrenamiento (es decir, 22 medidas). En definitiva, la red clasificadora fuzzy ARTMAP fue entrenada y validada 5 veces usando los 5 conjuntos de entrenamiento y validación y los porcentajes de acierto en la clasificación de los 5 bloques fue de 63.63 %, 95.45 %, 100% 100 % y 81.81 %. Promediando dichos resultados se puede hablar de un porcentaje promedio de acierto de 88.18 % en la clasificación, con una desviación estándar de 15.61 %. En total fueron 13 errores de las 109 medidas originales. Las confusiones ocurren entre las muestras que pertenecen a diferentes productores y la calidad de los diferentes jamones de un mismo productor. En particular también se produjeron confusiones entre muestras correspondientes a jamones obtenidos de cerdos que fueron alimentados con bellotas o con forraje.

El proceso de selección de variables se ejecutó usando los 5 conjuntos de entrenamiento y validación descritos anteriormente. Para cada par de conjuntos de entrenamiento y validación se realizó primero la selección de variables a través del conjunto de entrenamiento y con las variables seleccionadas se entrenó la red clasificadora fuzzy ARTMAP para luego determinar el porcentaje de aciertos de clasificación de los jamones utilizando el conjunto de validación. El primero y el segundo paso en la selección de variables se utilizaron para eliminar ruido y redundancia, tal y como se hizo en los conjuntos anteriores. Los valores umbrales empleados fueron de 0.5 y 0.8 para el criterio de la varianza y el de colinealidad respectivamente. Estos valores se determinaron analizando los resultados que se obtienen en los criterios de preselección para así, en base a los valores umbrales fijados seleccionar un número reducido de las variables. El número de variables que se mantuvieron después de estos dos pasos fue, en promedio, de 42.

Al igual que en los trabajos anteriormente descritos, el tercer paso en la selección de variables consistió en realizar una selección mas fina entre las variables restantes usando un proceso de simulated annealing. Dicho proceso se llevó a cabo independientemente a cada uno de los 5 bloques. El algoritmo SA fue ejecutado con 50 diferentes temperaturas y el número de iteraciones por temperatura fue de 40. Los

ISBN:978-84-693-4070-7/DL:T-1167-2010

clasificadores fuzzy ARTMAP (uno por bloque) fueron entrenados y validados usando como entradas las variables que se seleccionaron después de este ultimo paso. El porcentaje de aciertos en la clasificación de las muestras fue de 81.18 % 95.45%, 90.90% 100 % y 90.90 % con un porcentaje de acierto promedio de 91.68% en la clasificación de los jamones (la desviación estándar fue de 6.98%). Esto corresponde en promedio a 9 muestras mal clasificadas de las 109 totales. La confusión ocurrió entre medidas de diferentes productores pero nunca entre diferente calidades de jamones de un mismo productor. El número promedio de variables seleccionadas después de los tres pasos de selección fue de 14 (véase tabla 4.13), es decir solo un 8 % de las variables iniciales se mantuvieron. La tabla muestra que un alto número de variables son compartidas por los diferentes bloques de validación. Esto demuestra la robustez de los métodos de selección de variables aplicados.

Bloque #	Fragmentos seleccionados m/z
1	45, 47, 49, 56, 58, 59, 64, 70, 71, 73, 77, 79, 80, 81, 83, 85, 94, 100, 104, 111, 114
2	45, 47, 49, 53, 56, 57, 58, 60, 61, 64, 71, 72, 77, 81, 83, 84, 94, 100, 114, 208
3	47, 48, 55, 61, 64, 67, 77, 81, 82, 83, 84, 104, 138
4	45, 49, 56, 58, 64, 71, 77, 81, 82, 83, 84, 104, 138
5	45, 47, 51, 53, 56, 57, 58, 60, 61, 64, 67, 69, 70, 71, 72, 73, 79, 81, 82, 83, 84, 85, 93, 94, 101, 105, 108, 114, 133, 138
Fragmentos mas frecuentes m/z	45, 47, 49, 56, 58, 64, 71, 77, 81, 83, 84, 94, 114, 138

Tabla 4.13 fragmentos m/z seleccionados para cada bloque selección/validación después de los tres pasos de selección. La ultima fila muestra los fragmentos m/z más frecuentes.

Finalmente, se realizó la clasificación de las muestras en 11 categorías por medio de un clasificador fuzzy ARTMAP usando el resultado de los pasos anteriores de selección de variables. Solo las 14 relaciones m/z seleccionadas más frecuentemente se utilizaron como entradas del clasificador (véase tabla 4.13 para mas detalles), obteniendo un resultado en la clasificación de las 11 categorías de los jamones ibéricos del 94.54 % usando una validación cruzada de orden uno.

Sólo seis de las 109 muestras de jamón fueron clasificadas de forma incorrecta. Además, fue posible discriminar perfectamente entre jamones de cerdos alimentados con bellota y los alimentados con forraje ya que el resultado en la clasificación de estas dos categorías fue del 100 % de aciertos. Estos resultados son muy buenos comparados con el que se obtuvo empleando la totalidad de las variables disponibles (209) ya que fue del 88.18 % de acierto en la clasificación con la red fuzzy ARTMAP.

Los resultados obtenidos, en lo que concierne a fragmentos seleccionados por los algoritmos de selección de variables, pueden ser explicados aproximadamente con los niveles de ciertos compuestos químicos. Las diferencias en la alimentación del cerdo conducen a diversos perfiles volátiles presentes en los espacios de cabeza de muestras trituradas de la carne subcutánea de los cerdos. Los niveles de hexanol y pentanol, que aparecen principalmente por la oxidación del ácido linoleico, se presentan de forma similar sin importar la alimentación del cerdo. Por otra parte, el nonanal es el aldehido más importante derivado del ácido oleico que se encuentra en cantidades perceptiblemente mayores en los cerdos alimentados con bellotas que en los cerdos alimentados con forraje.

El fragmento m/z 114 seleccionado en el modelo está presente en el espectro de masas característico del nonanal. Por eso ayuda a discriminar entre los cerdos alimentados con bellota y los alimentados con forraje. Otros fragmentos seleccionados tales como el m/z 77 pueden presentarse en volátiles aromáticos; la m/z 71 indica presencia de esteres, alcanos, propilcetonas y butanoatos, y la m/z 45 podría ser debido a la presencia de ácidos carboxílicos o alcoholes. Finalmente, la presencia de pentilcetonas y metilcetonas son reveladas por las m/z 56 y 58,

ISBN:978-84-693-4070-7/DL:T-1167-2010

respectivamente. Todos estos compuestos han sido identificados como característicos del headspace de los jamones ibéricos curados en seco.

4.4 Selección de variables empleando support vector machines (SVM) para aplicaciones en sistemas de olfato electrónico.

4.4.1 Introducción

Como tercera parte de este capítulo, se muestra un nuevo procedimiento de selección de variables inspirado en una selección secuencial "backward" pero especialmente diseñada para trabajar con Support Vectors Machines (SVM). SVM es un método de reconocimiento de patrones no paramétrico con muchas propiedades que los hacen especialmente atractivo para los sistemas de olfato electrónico. En particular, son muy útiles para resolver problemas de clasificación y cuantificación. Los SVM realizan una minimización del riesgo estructural (es decir minimizan la dimensión del modelo), lo que resulta en una mayor habilidad de generalización. Además el método es computacionalmente sencillo, por lo que la aplicación de SVM es rápida (para una mayor información con respecto a este tipo de algoritmos consultar el capítulo tres).

Para evaluar la utilidad del proceso de selección de variables acoplado a los Support Vector Machines en problemas de clasificación y regresión se analizaron dos de los conjuntos tratados en trabajos anteriores: el primer conjunto corresponde a diferentes concentraciones sintéticas de vapores y sus mezclas binarias medidas con una nariz electrónica basada en sensores de gases de óxidos metálicos y el segundo conjunto correspondió a diferentes jamones ibéricos medidos con una nariz electrónica basada en espectrometría de masas [6].

4.4.2 Selección de variables y Support vector machines

El proceso de selección de variables utilizando SVMs fue desarrollado implementando un nuevo método de selección inspirado en la concatenación de varios procesos de "backward selection", proceso diseñado especialmente para trabajar en problemas de clasificación y regresión.

El procedimiento de selección de variables se realizó de la siguiente manera: en el primer paso, el cuadrado de la normal del vector que define el hiperplano óptimo para la separación (o regresión), $\|\omega_{s0}\|^2$, se calculó usando todas las variables disponibles. Una vez calculado, se estudió el efecto de eliminar la variable *i-ésima* en el hiperplano de decisión, calculando el factor de mérito como:

$$\delta_i = \left| \|\omega_{s0}\|^2 - \|\omega_{si}\|^2 \right| \quad \text{for } i = 1,\ldots,n. \tag{4.1}$$

donde $\|\omega_{si}\|^2$ es el resultado del hiperplano calculado sin la variable que no se ha tenido en cuenta. Dicho proceso se repite con cada una de las variables. Un valor δ elevado para una variable concreta implica que dicha variable es importante en la construcción del modelo, por lo que se debe mantener para el proceso de clasificación o regresión.

4.4.3 Selección de variables y clasificación usando SVM

Inicialmente se analizó el conjunto de medidas sintéticas intentando clasificar en seis categorías cada una de las medidas del conjunto. Estas categorías corresponden a la identificación de los tres vapores y sus tres mezclas binarias, independientemente de su concentración. Cada categoría quedó representada por 16 medidas (es decir 4 concentraciones diferentes por 4 repeticiones).

Un uno contra todos fue la estrategia empleada para construir el clasificador SVM. Esto implica que para identificar cada medida se construyeron 6 modelos

ISBN:978-84-693-4070-7/DL:T-1167-2010

SVM. El primer SVM fue entrenado con muestras de entrenamiento con valores positivos (1) para la clase 1 y todas las otras muestras de entrenamiento (las de otras categorías) con valor negativo (-1), es decir de forma binaria. Esto se repite para cada clase; por ejemplo, para la segunda clase se hace de forma similar, es decir, un valor de (1) para la clase 2 y (-1) para las otras clases, y así sucesivamente hasta tener los 6 modelos SVM relacionados a las 6 clases.

Para estimar la identificación de los vapores, se aplicó una estrategia de validación en bloque empleando los modelos SVM. Se crearon cuatro conjuntos diferentes de entrenamiento y validación de la siguiente forma: 1 medida por vapor o mezcla de vapor y concentración integran el conjunto de validación (es decir la primera replica de la medida fue usada en el primer conjunto de validación, la segunda medida se utilizó en el segundo conjunto de validación, y así sucesivamente). Las restantes 3 medidas por vapor o mezcla de vapores y concentración conforman el conjunto de entrenamiento. De esta forma 4 conjuntos de validación (24 medidas en cada conjunto), y cuatro conjuntos de entrenamiento (72 medidas en cada conjunto) fueron obtenidos.

Se probaron diferentes funciones kernel como son la lineal, la polinomial (de orden 2) y base radial. El mejor resultado se encontró cuando se usó como función kernel la polinomial de orden 2. El primer bloque de entrenamiento se empleó para determinar el valor óptimo del parámetro *C* usando una validación cruzada de orden uno. El mejor resultado se obtuvo cuando el valor de *C* es superior o igual a 2^{17}. Un valor alto para este parámetro implica que prácticamente cada muestra de entrenamiento se clasificó correctamente al finalizar el proceso de entrenamiento, ya que se penaliza enormemente cualquier error durante el entrenamiento.

En el primer paso, se construyeron y validaron 6 modelos de clasificación SVM usando las 120 variables de entrada disponibles. En el segundo paso, el proceso de selección de variables descrito anteriormente se empleó para reducir el número de variables de entrada. Por cada modelo SVM de clasificación, el proceso de selección de variables se aplicó cuatro veces usando los datos en los cuatro conjuntos de entrenamiento. Cada proceso de selección de variables y los SVM construidos fueron validados usando los correspondientes conjuntos de validación.

La tabla 4.14 muestra los resultados en la clasificación. Cuando se usaron las 120 variables, el porcentaje de acierto de identificación fue de 90.73 %. Este resultado se obtiene de multiplicar los diferentes porcentajes en la identificación de los 6 modelos SVM. El porcentaje de acierto particular de un modelo SVM se obtiene haciendo un promedio de aciertos en la identificación sobre los 4 bloques de la validación cruzada.

Clase #	bloque #	Porcentaje acierto clasificación (%)		# Variables
		Sin VS	Con VS	después VS
1	1	100	100	3
	2	100	100	3
	3	100	100	3
	4	100	100	3
2	1	100	100	8
	2	100	100	8
	3	100	100	8
	4	100	100	8
3	1	100	100	25
	2	100	100	23
	3	95.83	100	25
	4	100	100	26
4	1	95.83	95.83	26
	2	100	100	26
	3	100	100	20
	4	95.83	95.83	20
5	1	100	100	15
	2	95.83	95.83	15
	3	100	100	15
	4	100	100	20
6	1	95.83	95.83	26
	2	95.83	95.83	20
	3	91	91	26
	4	95.83	95.83	20

Tabla 4.14 Porcentajes de aciertos en la clasificación de los vapores y sus mezclas usando SVMs y el procedimiento uno contra todos con o sin selección de variables (SV).

ISBN:978-84-693-4070-7/DL:T-1167-2010

Cuando el proceso de selección se empleó el porcentaje de acierto total en la identificación fue de 91.72 %. Por lo tanto, el resultado en la clasificación no se altera pero la dimensionalidad de los patrones de entrada se redujo en un factor que se extendía entre 46 y 40, dependiendo del modelo de SVM.

Usando los 4 bloques de validación implica que el proceso de selección se implemente 4 veces para cada modelo de SVM. Esto podría conducir a seleccionar diferentes variables por cada uno de los 4 conjuntos de entrenamiento y validación. En la práctica no hay una alta variabilidad entre los parámetros seleccionados al construir un modelo dado. Por ejemplo, las variables seleccionadas al construir el primer modelo son siempre las mismas, lo que ocurre igual es el caso para el segundo modelo. Sin embargo, en los otros modelos existe una variabilidad en el número de variables seleccionadas con cada bloque (ver tabla 4.14).

Sin embargo, cuando sucede esto la mayoría de las variables se comparten entre los bloques. Por ejemplo, en el modelo 3, el número de variables fue de 23 (bloque 2) y 26 (bloque 4) pero las 23 variables seleccionadas en el bloque dos también aparecen en el bloque 1, 3 y 4.

Para el caso de los jamones ibéricos (segundo conjunto) se empleó un procedimiento similar al anterior. Inicialmente se consideró un problema de clasificación de 2 categorías, el cual consiste en discriminar jamones de cerdos alimentados con bellotas (clase 1) o con forraje (clase 2). En este caso 89 medidas (40 correspondientes a la clase 1 y 49 a la clase 2) se seleccionaron de forma aleatoria para integrar los conjuntos de selección de variables (o entrenamiento). Las 20 medidas restantes (10 corresponden a la clase 1 y las otras 10 a la clase 2) forman el conjunto de validación.

De nuevo, se implementó un método de validación cruzada usando el conjunto de entrenamiento para determinar el valor de C y elegir la función kernel. Seleccionando una función kernel polinomial de orden 2 y un valor de C de 2^7. Se construyó un modelo SVM empleando las 206 variables de entrada disponibles. El modelo se validó usando los patrones del conjunto de validación. El porcentaje de acierto en la identificación fue de un 60 %. Seguidamente se implementó el método

de selección de variables, de forma que sólo se seleccionaron 31 variables de las 206 totales. Las figura 4.6 esquematiza dicho proceso. Posteriormente se construyó un nuevo modelo SVM utilizando el conjunto de variables de entrada reducido. El porcentaje de acierto en la identificación fue del 100 %. En este caso, el procedimiento de selección reduce el número de variables de entrada en un factor de 67, y consecuentemente incrementa drásticamente la capacidad de generalización del clasificador SVM.

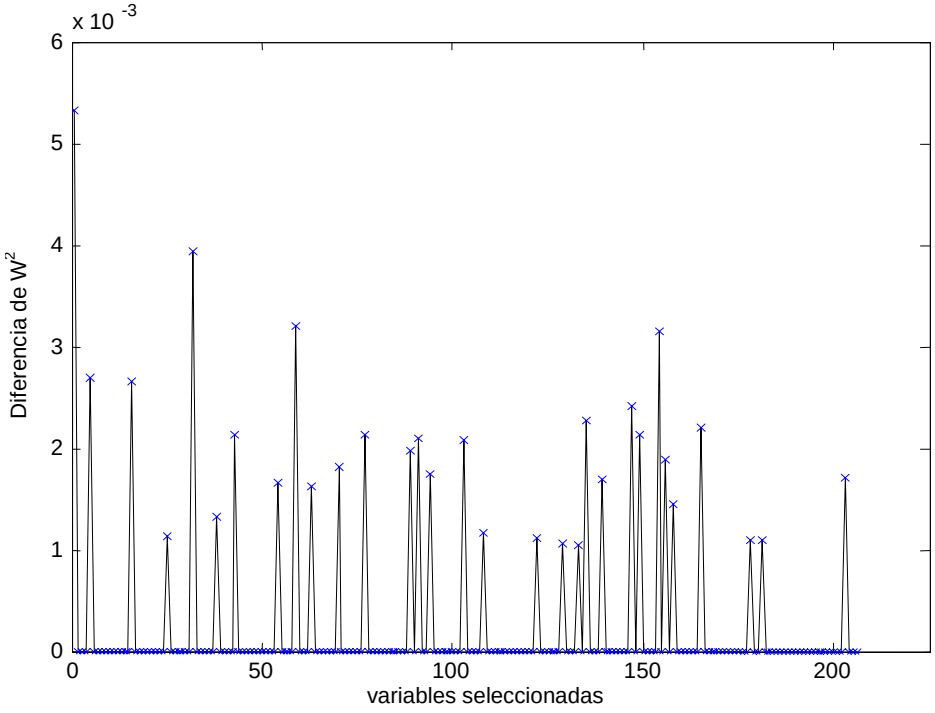

Figura 4.6 Clasificación en dos categorías. En el modelo SVM final se usan sólo 31 variables con un δ por encima de un umbral predeterminado.

En un segundo experimento se intentó entrenar un modelo SVM para clasificar las muestras de jamón en 11 categorías diferentes (ver tabla 4.15). En forma similar al estudio hecho a los vapores y sus mezclas se aplicó un uno contra todos (11 modelos SVM, es decir, se construyó uno por clase) y a su vez se implementaron 6 bloques de

ISBN:978-84-693-4070-7/DL:T-1167-2010

validación. Un conjunto de entrenamiento comprende 8 repeticiones (de las 10 disponibles) por tipo de jamón (es decir, 87 medidas en total, ya que sólo un tipo de jamón disponía de 9 medidas en vez de las 10 del resto).

Clase #	Porcentaje de aciertos clasificación (%)		# Variables después SV
	Sin SV	Con SV	
1	98.16	99.08	37
2	96.33	98.16	46
3	98.16	98.16	26
4	100	100	34
5	99.08	100	28
6	99.08	98.16	39
7	99.08	100	28
8	100	100	31
9	96.33	99.08	14
10	97.25	99.08	57
11	94.49	98.16	36
Total	79.91	90.30	-

Tabla 4.15 Porcentaje de aciertos en la clasificación de jamones ibéricos usando SVM empleando uno contra todos con y sin selección de variables (SV).

Los porcentajes de acierto para un modelo SVM en particular se obtiene promediando los porcentajes de identificación de los 6 bloques (este promedio se muestra en la tabla). Cuando el proceso de selección se empleó el porcentaje de acierto total promediando los resultados de identificación de las 11 clases fue de 90.30 %. Por consiguiente el resultado de la clasificación mejoró notablemente y,

adicionalmente, la dimensionalidad de los patrones de entrada se redujo en un factor que se extendía entre 3.7 y 14.9, dependiendo del modelo de SVM.

4.4.4 Selección de variables y regresión usando SVMs

En este estudio, el conjunto de jamones ibéricos fue utilizado para intentar realizar una regresión con SVM. Tres modelos específicos de regresión SVM se construyeron para predecir la humedad, la actividad del agua y contenido en sal en las muestras de jamón. El funcionamiento de los diferentes modelos se estimó mediante el cálculo de los coeficientes de correlación de la regresión lineal entre los valores reales y los resultados estimados mediante los modelos de predicción SVM. Solo las medidas de validación se usaron para calcular estos coeficientes de correlación. Una vez más, 6 bloques de validación se emplearon y los 6 conjuntos de validación y entrenamiento fueron los usados para el propósito de clasificación. Como en el análisis previo, el primer conjunto de entrenamiento se utilizó para determinar la función kernel y el valor de C. se utilizó una función kernel base radial con $\sigma = 3$ y $C = 2^{17}$ para los tres diferentes modelos de regresión. Una función ε-insensible se utilizo donde el valor de ε a utilizar se fijó en 0.1. Inicialmente, los modelos de regresión se construyeron y validaron usando las 206 variables de entrada disponibles.

En el segundo paso los modelos de regresión fueron construidos después de que el proceso de selección se realizara. El proceso de selección de variables fue equivalente al descrito en el análisis de la clasificación. Las 30 mejores variables (aquellas con valores altos de δ) se seleccionaron y fueron empleadas por los modelos de regresión SVM. La tabla 4.16 muestra los valores de los coeficientes de correlación de la regresión lineal entre los valores reales y predichos de humedad, actividad de agua y contenido en sal en las muestras de jamón. La tabla 4.16 también muestra que los diferentes modelos de regresión SVM funcionan de forma satisfactoria y que su funcionamiento incluso mejora cuando se llevó a cabo el proceso de selección de variables.

ISBN:978-84-693-4070-7/DL:T-1167-2010

Marca del jamón	Nombre corto	Tipos de jamón	Humedad (%)	Actividad de agua	Sal (%)	Cerdo alimentado
Extremadura	EX	EX_1 EX_2 EX_3 EX_4	36.46 41.83 36.97 32.33	0.745 0.797 0.776 0.720	7.39 7.03 6.68 6.03	bellota
Guijuelo #1	G1	G1	37.82	0.765	8.07	bellota
Huelva	HU	HU	45.91	0.807	7.64	bellota
Guijuelo #2	G2	G2_1 G2_2	41.17 45.28	0.849 0.800	3.65 8.60	forraje
Guijuelo #3	G3	G3_1 G3_2 G3_3	46.76 33.54 50.74	0.811 0.836 0.819	8.49 5.16 7.46	forraje

Tabla 4.16. *Los 11 tipos de jamón ibérico analizados. Los jamones difieren en productor, tipo de cerdo y forma de alimentación del cerdo.*

bloque #	Sin selección de variables			Con selección de variables		
Modelo #	1	2	3	1	2	3
1	0.855	0.858	0.780	0.947	0.964	0.955
2	0.964	0.938	0.908	0.974	0.985	0.983
3	0.955	0.915	0.884	0.975	0.979	0.976
4	0.963	0.950	0.942	0.988	0.969	0.945
5	0.951	0.962	0.933	0.987	0.960	0.860
6	0.933	0.919	0.920	0.976	0.972	0.933
Total	0.937	0.924	0.894	0.975	0.972	0.943

Tabla 4.17 Resultados de validación (coeficientes de correlación) de los modelos de regresión basados en SVM construidos para estimar humedad (modelo 1), actividad de agua (modelo 2) y sal (modelo 3) en muestras de jamón ibérico.

4.5 Conclusiones

Como conclusión general de este capítulo se puede decir que para los diferentes conjuntos estudiados, la inclusión de un proceso previo de selección de variables da como resultado una reducción drástica en la dimensionalidad de los datos y un aumento significativo en los correspondientes resultados de clasificación. Se ha demostrado que diferentes métodos (secuenciales o estocásticos) pueden ser acoplados a clasificadores fuzzy ARTMAP o PNN y ser usados para la selección de variables tanto en problemas de análisis de gases en sistemas multisensoriales como en sistemas olfativos basados en espectrometría de masas. Estos procesos permiten identificar y seleccionar un mínimo número de variables originales o fragmentos importantes para la discriminación correcta entre las mezclas en cada uno de los conjuntos estudiados. Por otro lado, se ha demostrado que implementando técnicas de reconocimiento de patrones basadas en SVMs en problemas de selección de variables se pueden obtener resultados prometedores en los procesos de clasificación o regresión en sistemas multisensoriales.

ISBN:978-84-693-4070-7/DL:T-1167-2010

4.6 Referencias

[1] Margalef Pedrola Pere Joan, “Proyecto final de carrera Avaluacio de xarxes ressonants en la clasificacio de mostres gasoses”. Universidad Rovira i Virgili . junio 2002.

[2] The Mathworks Inc., Matlab (versió 6.0), The Mathworks. Inc http:// www.mathworks

[3] Eduard Llobet., Jesús Brezmes, Oscar Gualdrón, Xavier Vilanova, Xavier Correig “Building parsimoniouus fuzzy ARTMAP models by variable selection with a cascaded genetic algorithm; application to multisensors systems for gas analysis” presentado en la revista Sensors and Actuators B. Vol 99, (2004) 267-272

[4] O. Gualdrón, E. Llobet, J. Brezmes, X. Vilanova, X. Correig “Coupling fast variable selection methods to neural network based classifiers: Application to multisensor systems.” Presentado en la revista Sensors and Actuators B, Vol 114, (2006) 522-529.

[5] E. Llobet; O. Gualdrón; M. Vinaixa; N. El-Barbri; J. Brezmes; X. Vilanova; B. Bouchikhi; R. Gómez; J.A. Carrasco; X. Correig. “Efficient feature selection for mass-spectrometry based electronic nose applications.” Presentado en la revista Chemometrics and Intelligent Laboratory Systems. (In Press).

[6] O. Gualdrón, J. Brezmes, E. Llobet, A. Amari, X. Vilanova, B. Bouchikhi, X. Correig “Variable selection for support vector machine based electronic noses” presentado en la revista Sensors and Actuators B. (In Press).

UNIVERSITAT ROVIRA I VIRGILI
DESARROLLO DE DIFERENTES MÉTODOS DE SELECCIÓN
DE VARIABLES PARA SISTEMAS MULTISENTORIALES
Oscar Eduardo Gualdron Guerrero
ISBN:978-84-693-4070-7/DL:T-1167-2010

5.

Conclusiones

ISBN:978-84-693-4070-7/DL:T-1167-2010

5.1 Conclusiones

Uno de los principales inconvenientes que en la actualidad presentan los sistemas de olfato artificial es la alta dimensionalidad de los datos obtenidos de las muestras analizadas, debido a la gran cantidad de parámetros que se obtienen de cada medida. El principal objetivo de esta tesis ha sido estudiar y desarrollar nuevos métodos de selección de variables con el fin de reducir la dimensionalidad de los datos y así poder optimizar los procesos de identificación, clasificación y/o cuantificación en sistemas de olfato electrónico basados en sensores de gases o en espectrometría de masas.

El tema de la selección de variables ha visto incrementado enormemente su interés en los últimos años, ya que la mayoría de los investigadores se han percatado de la importancia de identificar los parámetros clave (marcadores) inherentes a cada aplicación. De hecho, se puede afirmar sin ningún género de dudas, que este ha sido uno de los principales temas tratados en recientes congresos internacionales sobre sistemas de olfato electrónico como el ISOEN 2005, EuroSensors 2005, o el IEEE Sensors 2004 y 2005. Precisamente a estos congresos es a los que se han enviado los trabajos desarrollados en esta tesis, quedando demostrado que su temática está de plena actualidad en el campo de los sensores químicos y sistemas de olfato electrónico.

El inicio de las tareas englobadas en esta tesis doctoral consistió en la identificación de los principales métodos empleados por otros investigadores para tratar esta problemática. A partir de este estudio inicial se determinaron que métodos de selección de variables serian idóneos en aplicaciones de olfato electrónico, algoritmos que se acoplarían a modelos predictivos basados en diferentes redes neuronales como fuzzy ARTMAP y PNN, además de métodos de reconocimiento de patrones como los Support Vector Machines (SVM).

Para poder evaluar la importancia de los métodos y comprobar si ayudan realmente a solucionar la problemática de la elevada dimensionalidad se han utilizado cuatro conjuntos de datos pertenecientes a aplicaciones reales que nos

permitieron comprobar y comparar los diferentes métodos implementados de forma objetiva. Estos cuatro conjuntos de datos se han utilizado en tres estudios cuyas conclusiones repasamos a continuación:

En el primero de los estudios se ha demostrado que diferentes métodos (secuenciales o estocásticos) pueden ser acoplados a clasificadores fuzzy ARTMAP o PNN y ser usados para la selección de variables en problemas de análisis de gases en sistemas multisensoriales. Los métodos fueron aplicados simultáneamente para identificar y cuantificar tres compuestos orgánicos volátiles y sus mezclas binarias construyendo sus respectivos modelos neuronales de clasificación.

Los algoritmos de selección tales como el forward, backward y el stepwise pueden ser usados con éxito, pero cuando hay una gran cantidad de posibles soluciones pueden quedar atrapados en mínimos locales en el proceso de optimización. Por este motivo se optó por probar y comparar dicha metodología con algoritmos estocásticos como los algoritmos genéticos (GA) y el simulated annealing.

Aunque los GA y el simulated annealing pueden encontrar buenas soluciones en las aplicaciones consideradas en este conjunto sintético (casi siempre mejores que con métodos secuenciales) presentan el inconveniente de la gran carga computacional que requieren. Una posible solución a este dilema consiste en introducir un paso previo en el cual se calcula una figura de mérito para cada variable (relación intra varianza e inter varianza) que permite tener un criterio sólido para retener solamente aquellas variables que estén por encima de un umbral determinado (es decir, un determinado valor para su figura de mérito) para su procesado posterior. Este paso permite reducir el tiempo de cómputo aproximadamente en un factor de 4.

El segundo trabajo que se incluye en esta memoria propone una nueva estrategia para la selección de variables que se ha mostrado eficaz ante diferentes conjuntos de datos provenientes de sistemas olfativos basados en espectrometría de masas (MS). La selección de variables consistió básicamente en tres pasos. Los dos primeros tienen como principal objetivo eliminar información irrelevante y variables altamente colineales respectivamente. La eliminación de dichas variables tiene un coste computacional relativamente bajo que permite reducir drásticamente el número de

ISBN:978-84-693-4070-7/DL:T-1167-2010

variables (cerca del 80 % de las variables iniciales son eliminadas después del segundo paso).

Esto es muy interesante para sistemas de olfato electrónico basados en MS donde el número de parámetros (fragmentos m/z) disponible por medida es muy alto. Finalmente, en el tercer paso se aplica el algoritmo Simulated Annealing que realiza una selección mas fina sobre el conjunto reducido de variables que han sobrevivido a la selección de los pasos anteriores.

La estrategia ha sido aplicada inicialmente a un conjunto de datos consistente de mezclas sintéticas de compuestos volátiles. Este conjunto ha sido usado para mostrar que el proceso de selección es viable para identificar un mínimo número de fragmentos que permiten la discriminación correcta entre mezclas usando clasificadores fuzzy ARTMAP. Además, dada la naturaleza simple del problema planteado, fue posible mostrar que los fragmentos seleccionados, son fragmentos de ionización característicos de las especies presentes en las mezclas a ser discriminadas. Una vez demostrado el correcto funcionamiento de esta estrategia, se aplicó esta metodología a otros dos conjuntos de datos (aceite de oliva y jamones ibéricos, respectivamente).

Cuando el método se aplicó al conjunto de aceites de oliva, el número de variables pudo reducirse de 106 a 4. Usando solo 4 fragmentos m/z y un clasificador fuzzy ARTMAP fue posible identificar correctamente las muestras según su variedad de oliva y su origen con un 100 % de acierto, un resultado significativamente mejor al obtenido con la totalidad de los fragmentos con el que se consiguió un 94.44 % de acierto en la clasificación. Este resultado demuestra que la determinación, automática o manual, de las variables clave del proceso es fundamental para llevar a buen término los resultados de la aplicación estudiada El proceso completo de selección de las variables empleó un tiempo total de 20 minutos aproximadamente en una plataforma PC Pentium 4.

Aplicando el método al conjunto de datos de los jamones ibéricos se redujo el número de variables de 209 a 14. Usando las variables supervivientes, un clasificador fuzzy ARTMAP fue capaz de discriminar muestras de jamón de acuerdo al productor y a su calidad (clasificación de las 11 categorías), con un 97.24% de acierto.

Además, fue posible identificar con un 100 % de acierto si los cerdos habían sido alimentados con bellota o forraje. Comparando estos resultados con los obtenidos utilizando la totalidad de las variables donde el porcentaje de acierto promedio fue de 88.18% en la clasificación, podemos deducir que los resultados son significativamente mejores.

Como conclusión se puede decir que para los diferentes conjuntos estudiados, la inclusión de un proceso previo de selección de variables da como resultado una reducción drástica en la dimensionalidad y un aumento significativo en los correspondientes resultados de clasificación. Los métodos introducidos aquí no solo son útiles para resolver problemas de narices electrónicas basadas en MS, sino también para cualquier aplicación de sistemas de olfato artificial que presenten problemas de alta dimensionalidad como en el caso de los conjuntos de datos estudiados en este trabajo.

El tercer estudio tratado en esta tesis ha girado en torno al desarrollo de un nuevo método de selección de variables inspirado en la concatenación de varios procesos de "backward selection". El método está especialmente diseñado para trabajar con SVMs en problemas de clasificación o de regresión.

La utilidad del método ha sido evaluada usando dos de los conjuntos de datos ya utilizados anteriormente. El primer conjunto utiliza medidas sintéticas de vapores y sus mezclas usando una matriz de sensores y el segundo conjunto fue el de las medidas de jamones ibéricos.

Se construyó un modelo de clasificación SVM usando todas las variables de entrada y posteriormente también usando un pequeño conjunto de variables seleccionadas mediante un novedoso algoritmo basado en el peso relativo de cada variable en el modelo SVM. Aplicar este proceso de selección dio como resultado una reducción drástica en el número de variables de entrada usadas por el modelo SVM y un significativo incremento en la capacidad de clasificación. Así, se paso de un 79.91 % de acierto usando todas las variables a un 90.3% usando un conjunto reducido de variables en la clasificación de las 11 categorías de jamones ibéricos, alcanzando un 100% de clasificación correcta de jamones de cerdos alimentados con

ISBN:978-84-693-4070-7/DL:T-1167-2010

bellotas o con forraje. Este rendimiento fue estimado aplicando un riguroso método de validación por bloques.

Finalmente hay que destacar que el método de selección de variables ayudó también a mejorar el funcionamiento de los modelos para regresión. Concretamente, en el conjunto de datos de las muestras de jamón se obtuvo una buena exactitud en la predicción de la humedad, actividad del agua y el contenido de sal. Ambos procesos, la selección de variables y el reconocimiento de patrones mediante SVMs son computacionalmente sencillos, por lo que ambos métodos son de gran interés para un amplio espectro de tecnologías de sensado que se pueden incorporar en sistemas de olfato artificial.

Teniendo en cuenta que los conjuntos de datos analizados mediante SVMs son los mismos que fueron utilizados en estudios anteriores podríamos decir que los resultados en la clasificación empleando los diferentes métodos de selección de variables acoplados a redes neuronales y los obtenidos utilizando los SVM como método de selección y clasificación son muy similares, aunque el número de variables en el primer caso es mucho más reducido con respecto al segundo caso. Por otra parte es importante decir que el proceso utilizando SVM se realiza de forma binaria (es decir clasificación en dos clases o categorías) por lo que es necesario realizar el proceso tantas veces como clases se dispongan y no un análisis con la totalidad de las categorías en un solo bloque.

Como conclusión final de todo el trabajo realizado bajo el paraguas de esta tesis doctoral podemos afirmar que la selección de variables es un paso previo fundamental si se quiere tener éxito en el desarrollo de aplicaciones basadas en sistemas de olfato electrónico. Además de reducir la carga computacional del procesado de datos, este paso permite incrementar el nivel de acierto de este tipo de sistemas en aplicaciones reales. Además, la selección automática de variables clave en cualquiera de las aplicaciones estudiadas ha permitido conocer en mayor profundidad el problema químico al cual nos enfrentábamos, lo cual puede dar a luz soluciones o deducciones no contempladas a priori.

Finalmente cabe comentar que la selección de variables es un amplio campo que todavía puede experimentar mejoras interesantes y cuyas aplicaciones van más allá

de las estudiadas en este trabajo. Uno de los campos más prometedores para esta disciplina es en las denominadas ciencias “ómicas” (metabolómica, proteómica, genómica, etc), en las que la selección de variables se puede traducir directamente en la detección de bio-marcadores de gran interés en el campo de la medicina.

UNIVERSITAT ROVIRA I VIRGILI
DESARROLLO DE DIFERENTES MÉTODOS DE SELECCIÓN
DE VARIABLES PARA SISTEMAS MULTISENTORIALES
Oscar Eduardo Gualdron Guerrero
ISBN:978-84-693-4070-7/DL:T-1167-2010

UNIVERSITAT ROVIRA I VIRGILI
DESARROLLO DE DIFERENTES MÉTODOS DE SELECCIÓN
DE VARIABLES PARA SISTEMAS MULTISENTORIALES
Oscar Eduardo Gualdron Guerrero
ISBN:978-84-693-4070-7/DL:T-1167-2010

6.

Anexo: Lista de Publicaciones

ISBN:978-84-693-4070-7/DL:T-1167-2010

6.1 Publicaciones derivadas de esta tesis doctoral

[1] Eduard Llobet., Jesús Brezmes, Oscar Gualdrón, Xavier Vilanova, Xavier Correig "Building parsimoniouus fuzzy ARTMAP models by variable selection with a cascaded genetic algorithm; application to multisensors systems for gas analysis" presentado en la revista Sensors and Actuators B. Vol 99, (2004) 267-272

[2] O. Gualdrón, E. Llobet, J. Brezmes, X. Vilanova, X. Correig "Coupling fast variable selection methods to neural network based classifiers: Application to multisensor systems." " presentado en la revista Sensors and Actuators B, Vol 114, (2006) 522-529.

[3] E. Llobet; O. Gualdrón; M. Vinaixa; N. El-Barbri; J. Brezmes; X. Vilanova; B. Bouchikhi; R. Gómez; J.A. Carrasco; X. Correig. "Efficient feature selection for mass-spectrometry based electronic nose applications." Presentado en la revista Chemometrics and Intelligent Laboratory Systems. (In Press).

[4] O. Gualdrón, J. Brezmes, E. Llobet, A. Amari, X. Vilanova, B. Bouchikhi, X. Correig "Variable selection for support vector machine based electronic noses" presentado en la revista Sensors and Actuators B. (in Press).

6.2 Conferencias.

[1] O. Gualdrón. “Variable selection methods for e-nose applications” presentado en el PhD student’s Meeting on electron Devices and Microelectronics. Realizado en la Universidad Rovira i Virgili Tarragona España. Junio 2003

[2] O. Gualdrón, E. Llobet, J. Brezmes, X. Vilanova, X. Correig “Fast variable selection for gas sensing applications” presentado en IEEE Sensors 2004,Viena, Austria, Octubre 2004.

[3] O. Gualdrón, E. Llobet, J. Brezmes, X. Vilanova, X. Correig “Variable selection for support vector machines based pattern recognition” presentado en EuroSensors 2005,Barcelona, España 2005.

[4] E.Llobet, O. Gualdrón, J. Brezmes, X. Vilanova, X. Correig “An unsupervised dimensionality-reduction technique” presentado en IEEE Sensors 2005, California, EEUU, 2005.

[5] C. Duran, J.Brezmes, O.Gualdrón, M.Vinaixa, E.Llobet, X.Vilanova, X.Correig “Concatenation of a Fuzzy Artmap neural network to different variable selection techniques to enhance E-nose performance” presentado en ISOEN 2005, Barcelona, España 2005.

[6] O.Gualdrón. “Variable selection for Support Vector Machines based pattern recognition in multisensor systems” presentado en el PhD student’s Meeting on electron Devices and Microelectronics. Realizado en la Universidad Rovira i Virgili Tarragona España. Junio 2006.