Lat. Am. J. Aquat. Res., 40(1): 90-101, 2012
DOI: 10.3856/vol40-issue1-fulltext-9

Research Article

Clasificación acústica de anchoveta *(Engraulis ringens)* y sardina común *(Strangomera bentincki)* mediante máquinas de vectores soporte en la zona centro-sur de Chile: efecto de la calibración de los parámetros en la matriz de confusión

Hugo Robotham[1], Paul Bosch[1], Jorge Castillo[2] & Ignacio Tapia[1]
[1]Facultad de Ingeniería, Instituto de Ciencias Básicas, Universidad Diego Portales
P.O. Box 3249, Santiago, Chile
[2]Departamento de Evaluaciones Directas, Instituto de Fomento Pesquero
P.O. Box 8V, Valparaíso, Chile

RESUMEN. Se clasificó la anchoveta (*Engraulis ringens*) y sardina común (*Strangomera bentincki*) detectadas mediante equipos acústicos en la zona centro-sur de Chile, mediante el método de Máquinas de Vectores Soporte (SVM). Para esto se utilizaron descriptores de cardúmenes extraídos desde ecogramas, que fueron clasificados como morfológicos, batimétricos, energéticos y posicional espacial. Para lograr clasificaciones precisas mediante la utilización de esta metodología, fue necesario optimizar parámetros correspondientes al Kernel-Gaussiano, γ y de penalización del modelo C, mediante el análisis del efecto de la calibración sobre las matrices de confusión resultantes de la clasificación de las especies analizadas. El método SVM ajustó correctamente el 95,3% de los cardúmenes de anchoveta y sardina común. Los parámetros óptimos del Kernel-Gaussiano γ y de penalización C obtenidos mediante la metodología propuesta fueron γ = 450 y C = 0,95, respectivamente. Los parámetros mencionados incidieron de manera importante en la matriz de confusión y los porcentajes de clasificación final, por lo que se sugiere establecer, en aplicaciones futuras de este método, un protocolo experimental de calibración. La sardina común fue la especie con menor error de clasificación en el conjunto de las matrices de confusión. El descriptor correspondiente a profundidad del fondo fue el más sensible al SVM, la segunda variable en importancia es el descriptor distancia a la costa.

Palabras clave: máquinas de vectores soporte, clasificación de especies, hidroacústica, peces pelágicos, anchoveta, sardina, Chile.

Acoustic classification of anchovy *(Engraulis ringens)* and sardine *(Strangomera bentincki)* using support vector machines in central-southern Chile: effect of parameter calibration on the confusion matrix

ABSTRACT. The support vector machines (SVM) method was used to classify the anchovy (*Engraulis ringens*) and common sardine (*Strangomera bentincki*) species detected in south–central Chile by means of acoustic equipment. For this, descriptors of fish schools (morphology, bathymetry, energy, spatial position) extracted from ecograms were used. In order to obtain precise classifications using this methodology, it was necessary to optimize the parameters Gaussian-Kernel γ and penalty term C by analyzing the effect of the calibration on the confusion matrices resulting from the classification of the species under study. The SVM method correctly classified 95.3% of anchovy and sardine schools. The optimal parameters of the Gaussian-Kernel γ and penalty C obtained with the proposed methodology were γ = 450 and C = 0.95. These parameters have an important influence over the confusion matrix and the final classifications percentages, suggesting the development of experimental protocols for calibrating these parameters in future applications of this methodology. In all the confusion matrices, the common sardine showed the lowest classification error. The bottom depth was the descriptor that was most sensitive to the SVM, followed by school-shore distance.

Keywords: support vector machines, species classification, hydroacoustics, pelagic fishes, anchovy, sardine, Chile.

Corresponding author: Hugo Robotham (hugo.robotham@udp.cl)

INTRODUCCIÓN

En los últimos años las técnicas hidroacústicas han aumentado su popularidad para evaluar la abundancia y particularmente estudiar el comportamiento de varias especies de peces. Esto se debe a la alta resolución espacio-temporal, caracterizada por un muestreo no invasivo, que no altera a los organismos en su medio ambiente. Por otra parte, los avances tecnológicos, tales como el desarrollo de la electrónica y las capacidades de procesamiento de los computadores, han permitido en las últimas décadas realizar importantes progresos en la resolución de los equipos acústicos y, además, obtener mayores volúmenes de información. Sin embargo, la identificación acústica de los blancos aún es un motivo de investigación puesto que los equipos acústicos poseen baja resolución específica, es decir, pueden detectar blancos acústicos, pero no clasificarlos automáticamente por especies. Esta limitación puede inducir importantes sesgos en las estimaciones de abundancia a partir de esta metodología.

Algunos autores han utilizado diferentes técnicas con resultados favorables para la identificación automática de cardúmenes monoespecíficos (Horne, 2000; Fernandes *et al*., 2006), utilizando información de tamaño, localización y eco-intensidad de los cardúmenes de peces. Sin embargo, el problema persiste cuando se presentan cardúmenes multiespecíficos (Fernandes, 2009).

Generalmente, la clasificación de especies requiere de un proceso de escrutinio de los ecogramas, donde se combina el criterio del experto complementado con la información adicional que se adquiere de los lances de pesca (Simmonds & MacLennan, 2005). Este procedimiento incorpora un componente de arbitrariedad e incertidumbre, especialmente cuando se estudian poblaciones multiespecíficas de peces. En estos casos los resultados finales contarán con una cierta subjetividad la cual dependerá de la experiencia del operador encargado de realizar la evaluación de la especie objetivo.

Esta limitante, ha derivado en el desarrollo de otras aproximaciones alternativas o auxiliares de identificación de especies que ayuden a disminuir estos sesgos. Una de estas aproximaciones es la identificación de peces basadas en "parámetros descriptores acústicos" extraídos de los cardúmenes a partir de datos obtenidos a una única frecuencia acústica. Una segunda y cada vez más usual aproximación se basa en datos acústicos que consideran múltiples frecuencias (Korneliussen *et al*., 2009) combinadas con información geográfica y algunos parámetros morfológicos. La clasificación de cardúmenes monoespecíficos con información de una única frecuencia acústica se ha realizado mediante un amplio rango de técnicas estadísticas, destacándose los métodos multivariados como análisis de componentes principales y de funciones discriminantes (Nero & Magnuson, 1989; Vray *et al*., 1990; Scalabrin *et al*., 1996; Lawson *et al*., 2001). También se han aplicado métodos de clasificación como redes neuronales artificiales (Haralabous & Georgakarakos, 1996; Simmonds *et al*., 1996; Cabreira *et al*., 2009); análisis del vecino más cercano (Richards *et al*., 1991); conglomerados k-medias (Tegowski *et al*., 2003); modelos mixtos (Fleischman & Burwen, 2003), método de Kernel (Buelens *et al*., 2009); métodos árboles de clasificación (Fernandes, 2009). Demer *et al*. (2009) utilizan el análisis estadístico espectral. Fablet *et al*. (2009) aplican un modelo probabilístico. Korneliussen *et al*. (2009) combinan datos acústicos de multifrecuencia con información morfológica de los cardúmenes y distribución geográfica de las especies. Más recientemente, Robotham *et al*. (2010) usan Máquinas de Vectores Soporte (SVM) para clasificar cardúmenes de especies pelágicas en la costa de Chile y la compara con redes neuronales artificiales.

La técnica de SVM fue introducida por Vladimir Vapnik (Boser, 1992). Este método consiste en un conjunto de algoritmos de aprendizaje supervisado que resuelven problemas de clasificación y regresión. Dado un conjunto de ejemplos de entrenamiento (de muestras), se pueden etiquetar las clases y entrenar una SVM para construir un modelo que prediga la clase de una nueva muestra. Desde el punto de vista de la estructura y funcionamiento, los métodos de clasificación SVM son una potente herramienta para la solución de problemas de clasificación con gran cantidad de datos y atributos. La mayor dificultad del método recae en la calibración de algunos parámetros asociado al Kernel utilizado y el parámetro de penalización del modelo.

En este trabajo se aplicó el método SVM para clasificar a las especies anchoveta (*Engraulis ringens*)

y sardina común (*Strangomera bentincki*) detectadas acústicamente en la zona centro-sur de Chile a partir de descriptores de morfología, batimetría, energía acústica y posición espacial extraídos de los cardúmenes. Se analizó especialmente el efecto de la calibración de los parámetros correspondiente al Kernel Gaussiano, y de penalización del modelo C, sobre las matrices de confusión resultantes de la clasificación de las dos especies tenidas en cuenta en este estudio.

MATERIALES Y MÉTODOS

Obtención de los datos

La base de datos completa consideró cardúmenes provenientes de dos cruceros de evaluación por métodos acústicos utilizando el B/C Abate Molina realizadas en la zona norte y zona centro-sur de Chile (25°50'S a 41°00'S) en el año 2006 (Fig. 1). Esta base de datos fue depurada de acuerdo a criterios de cohabitación de especies; certeza en la identificación de especies mediante lances de pesca y condición monoespecífica de los cardúmenes. Se utilizó un ecosonda científico SIMRAD EK-500 operando un transductor de haz dividido ES38 de 38 kHz de frecuencia, el cual fue calibrado de acuerdo a procedimientos estándar (Foote *et al*., 1987). Los registros acústicos de los cardúmenes detectados por el ecosonda se procesaron mediante la utilización del programa de pos-procesamiento SonarData Echoview v.3.0. La identificación de los cardúmenes se efectuó mediante lances de pesca con una red de arrastre a media agua siguiendo el mismo plan de navegación. Los parámetros de los cardúmenes fueron extraídos automáticamente por el algoritmo SHAPES que se encuentra en el módulo denominado "schools" descrito por Barange (1994), Coetzee (2000) y Lawson *et al*. (2001) que permite identificar la silueta de las distintas agregaciones de organismos. Cada agregación fue marcada manualmente en una región sobre la imagen del ecograma y cada cardumen fue individualmente analizado. Los criterios para extraer los cardúmenes desde el ecograma fueron: altura y largo mínimo = 1 m; máxima distancia de enlace vertical = 1 m y máxima distancia de enlace horizontal = 15 m. Además, cada cardumen fue corregido de las distorsiones debidas a los efectos de la duración del pulso y el ancho del haz de sonido.

Análisis de datos

Los datos de entrada para clasificar los cardúmenes de peces constituyen una colección de registros acústicos. Cada registro acústico se caracterizó por un vector (x, y) donde "x" es el conjunto de descriptores acústicos, y donde "y" es el conjunto de categorías o de especies, como en este caso. Se usó un total de 12 descriptores por cardumen, los cuales fueron agrupados en distintas categorías como: morfológicas (altura media del cardumen, largo, perímetro, área, elongación y dimensión fractal); batimétricas (profundidad del fondo, profundidad media del cardumen, índice de altitud del cardumen); energía (energía acústica, densidad acústica) y de posición espacial (distancia a la costa del cardumen) (Scalabrín, 1991; Scalabrín & Massé, 1993). Algunos descriptores batimétricos y morfológicos se presentan en la Figura 2.

La técnica de clasificación SVM fue aplicada a los registros acústicos monoespecíficos plenamente identificados mediante lances de pesca cuya captura estuvo constituida por más del 90% de una sola especie. Además, fueron considerados solamente registros diurnos durante la estación de verano.

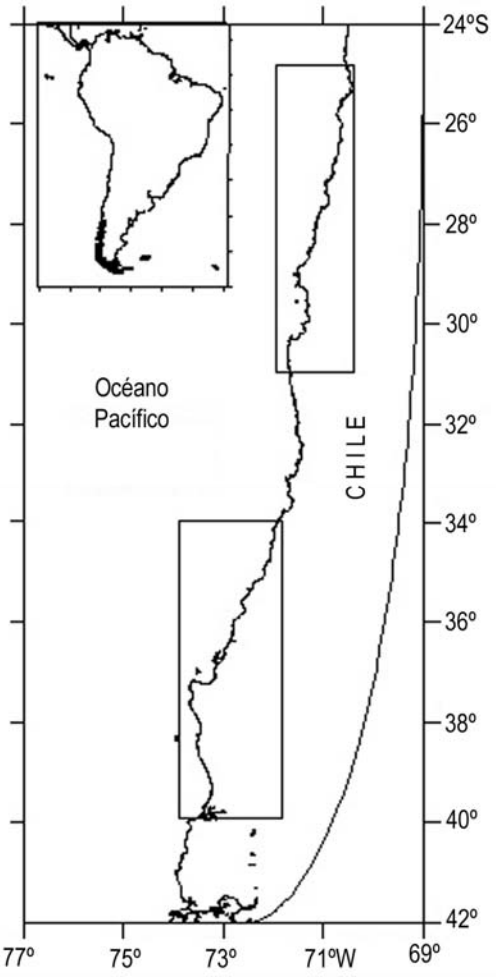

Figura 1. Área de evaluación hidroacústica de sardina común y anchoveta en 2006.

Figure 1. Hydroacustic surveys area of common sardine and anchovy in 2006.

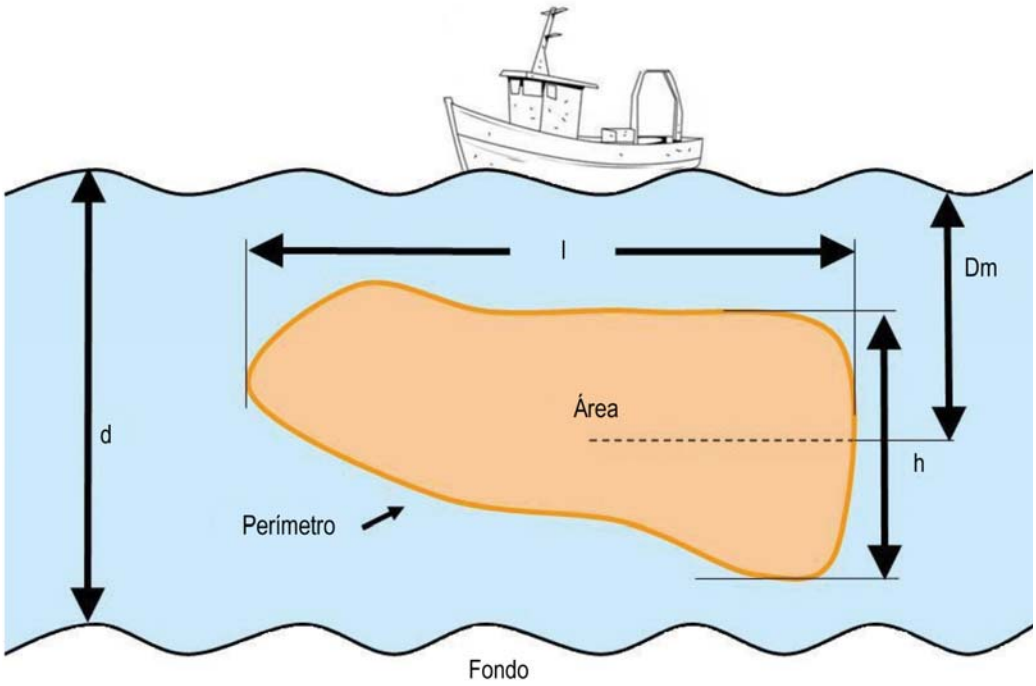

Figura 2. Algunos descriptores morfológicos y batimétricos en la columna de agua.
Figure 2. Some morphological and bathymetry descriptors in the water column.

Para conocer la contribución individual de cada descriptor en la clasificación se realizó un análisis de sensibilidad que consistió en eliminar un predictor por vez y evaluar el efecto en el error de salida. El valor del índice de sensibilidad se obtuvo como el cociente entre error del modelo al omitir un descriptor a la vez y el error del modelo con la totalidad de las variables. Un índice igual o cercano a 1 indica que el predictor tiene bajo peso en la estructura general del modelo.

Técnica de clasificación SVM

El SVM es una técnica de clasificación que ha recibido bastante atención. Esta técnica ha sido propuesta por Vapnik (1995) y pertenece a la familia de clasificadores lineales dado que busca separar el espacio de las características mediante hiperplanos. El SVM fue originalmente diseñado para clasificación binaria, sin embargo, este se puede extender a problemas multiclase (Weston & Watkins, 1998; Hsu & Lin, 2002).

Para entender mejor, supongamos que tenemos un conjunto de datos de dos tipos, llamados datos de entrenamiento: $\{(\chi_1, \gamma_1), (\chi_2, \gamma_2), \ldots, (\chi_n, \gamma_n)\}$ donde $\chi_i \in R^d$ representan el vector de predictores o de características de los datos y $\gamma_i \in \{-1,+1\}$ es la variable que permite diferenciar a cada especie, por ejemplo, en el presente caso permite discernir entre una anchoveta y una sardina. Un hiperplano de separación es una aplicación lineal en el espacio de las características, con coeficientes apropiado y que serán las variables del problema, y que permitirá la separación de los datos a partir del signo de esta aplicación cuando se evalúa en cada dato. La idea es encontrar la distancia mínima desde el hiperplano de separación hasta los datos más cercanos, esta distancia se denomina "margen" (Figura 3, tomada de Hastie *et al*., 2001). Un hiperplano de separación se llamará "optimal", si el margen (λ) es de tamaño máximo. De manera intuitiva, se ve claramente que un margen más grande corresponde a una mejor generalización. Por tanto, el problema de encontrar el hiperplano optimal es equivalente a encontrar $\beta \in R^d$ que maximice el margen. Los puntos de los datos que definen el margen a cada lado del hiperplano de separación se denominan "vectores soporte".

En muchos casos, no existe una separación lineal del problema, entonces mediante una transformación se lleva el espacio de los vectores predictores, cuya dimensión está determinada por la cantidad de predictores considerados en los datos, a un nuevo espacio de dimensión más grande y donde sí se puede encontrar un separador lineal, previa inclusión de un parámetro C, el cual permite cierto margen de error. Después de algunas transformaciones, el modelo inicial se generaliza para casos no separables a través de una transformación o función Kernel que debemos

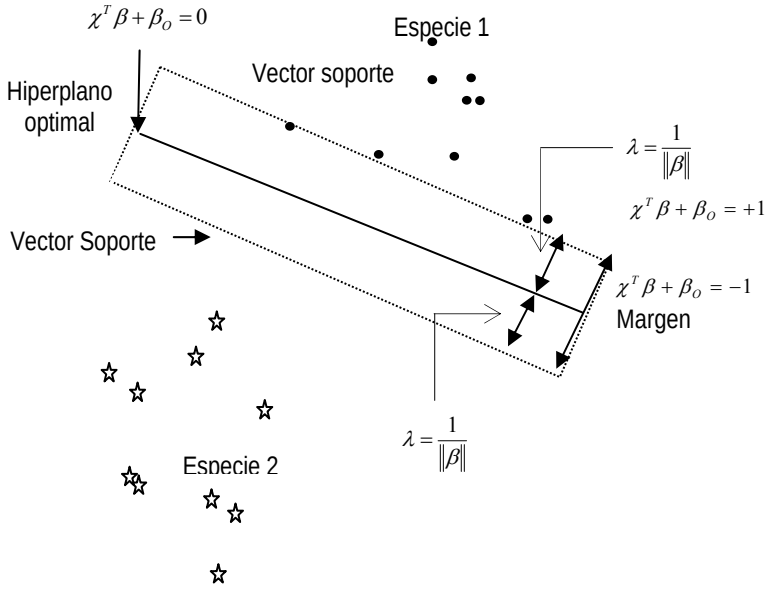

Figura 3. Hiperplano optimal del clasificador de vectores soporte.
Figure 3. Optimal hiperplane on support vector classifiers.

fijar (Kernel *K*) y que refleja el conocimiento de la naturaleza del problema. Sin embargo, para la clasificación, la literatura (Scholkopf & Smola, 2002) recomienda el uso del Kernel-Gaussiano. La recomendación se justifica por la estabilidad que muestra este Kernel, así como por el hecho práctico de que solamente se necesite estimar un único parámetro asociado a esta función el cual define el ancho que tiene el producto interno del Kernel. Lo segundo a determinar es el parámetro de regularización *C*, éste es independiente del Kernel que se esté considerando y representa un balance entre el tamaño del margen y el error de entrenamiento. Finalmente, la función de decisión viene dada por:

$$H(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i^{*} y_i K(x_i, \mathbf{x})$$

la cual clasifica a una observación dada en una especie u otra, en dependencia del signo de esta función evaluada en el vector de predictores. Los parámetros son la solución del problema de optimización cuadrático que modela esta metodología y *K* representa la función Kernel.

Muestra de entrenamiento y validación

Un subconjunto de 856 cardúmenes fueron seleccionados para el análisis de los patrones de reconocimiento: 442 de sardina común y 414 de anchovetas. La base de datos utilizada fue dividida en dos conjuntos: los datos de entrenamiento y los datos de validación del modelo. El proceso de elección de los datos se realizó de manera aleatoria por el programa MATLAB. Los resultados se obtuvieron con el programa computacional SVM light desarrollado por Joachims (2001).

De los 856 cardúmenes de la muestra, se seleccionaron 214 (25%, 107 de sardinas comunes y 107 de anchovetas), para ser utilizados en la prueba final de validación. Estos datos se mantuvieron aislados del resto y no participaron del proceso de entrenamiento, y solo se usaron en la etapa final del estudio. Los 642 cardúmenes restantes (75% del total) fueron utilizados para el proceso de entrenamiento, que correspondieron a 335 muestras de sardina común y 307 de anchoveta. Tanto las muestras de entrenamiento como validación pueden considerarse balanceadas, lo que representa una ventaja para el proceso de aprendizaje de los métodos heurísticos.

Cada descriptor usado en el experimento fue estandarizado con una media cero y varianza uno, para eliminar las diferencias de escala o magnitudes que puedan causar efectos no deseados en el modelo.

Experimento de calibración de parámetros

La calibración de la SVM consistió en determinar el valor de dos parámetros: el correspondiente al Kernel-Gaussiano, $\gamma = 1/\sigma$ y el parámetro de penalización del modelo C. El proceso de estimación del parámetro C consistió en probar combinaciones de parámetros en el intervalo (50, 950) con un tamaño de paso constante de 50 y entre (0,05; 0,95) para el Kernel γ con tamaño de paso de 0,05.

Luego se procedió a refinar el intervalo en una vecindad en torno a cada parámetro. Para cada prueba se realizaron 200 iteraciones. En cada una de ellas se procedió a dividir, de manera aleatoria, la muestra de 642 cardúmenes en 482 cardúmenes para crear la SVM y 160 para comprobarla. Estos cardúmenes mencionados corresponden, respectivamente al 75% y 25% de la muestra de entrenamiento (Robotham, 2010). En la Figura 4 se muestra un esquema que grafica este procedimiento. Una vez calibrado los parámetros y determinado el SVM se obtiene sobre la muestra de validación la matriz de clasificación denominada también matriz de confusión.

RESULTADOS

En la Tabla 1 se presenta un conjunto de cinco estadísticos descriptivos para los 12 descriptores de cardúmenes anchoveta y sardina común. En ambas especies, los indicadores morfológicos, altura media del cardumen y dimensión fractal presentaron un bajo coeficiente de variación (CV), respecto de los otros cuatro indicadores (largo, perímetro, área y elongación). La anchoveta presentó para estos cuatro indicadores morfológicos una media inferior a la de sardina común y a la vez, presentó una dispersión relativa (CV) más alta que el resto de los descriptores. Los estadísticos del descriptor batimétrico profundidad del fondo mostró importantes diferencias entre las dos especies, donde la anchoveta se asoció a una mayor amplitud de profundidad que la sardina común, esta última no sobrepasó los 328 m. Los tres indicadores batimétricos presentaron bajos CV respecto a los morfológicos y con órdenes de magnitud similares entre las dos especies. Los descriptores energéticos, energía acústica y densidad acústica de la sardina presentaron una media por cardúmen más alta que la anchoveta, lo mismo ocurrió con el CV del descriptor energía acústica. La distancia media a la costa presentó un CV bajo en ambas especies y posicionó a las anchovetas a una distancia media de 7,08 mn de la costa, casi el doble de las sardinas comunes 3,94 mn. A mayor distanciamiento de la costa mayor es la profundidad del fondo.

Es interesante hacer notar las diferencias en las características batimétricas de las dos zonas de estudio, mientras en la zona norte, la batimetría es profunda y aumenta abruptamente con la distancia a la costa, en la zona sur es más somera y las especies estudiadas se encuentran dentro de la influencia de la plataforma continental.

Por otro lado, el índice de altitud, relacionó la localización de los cardúmenes en la columna de agua respecto a la profundidad del fondo del mar. De este modo se determinó detectar que la anchoveta se encontraba más vinculada con la superficie, mientras que la sardina estuvo más relacionada con el fondo del mar.

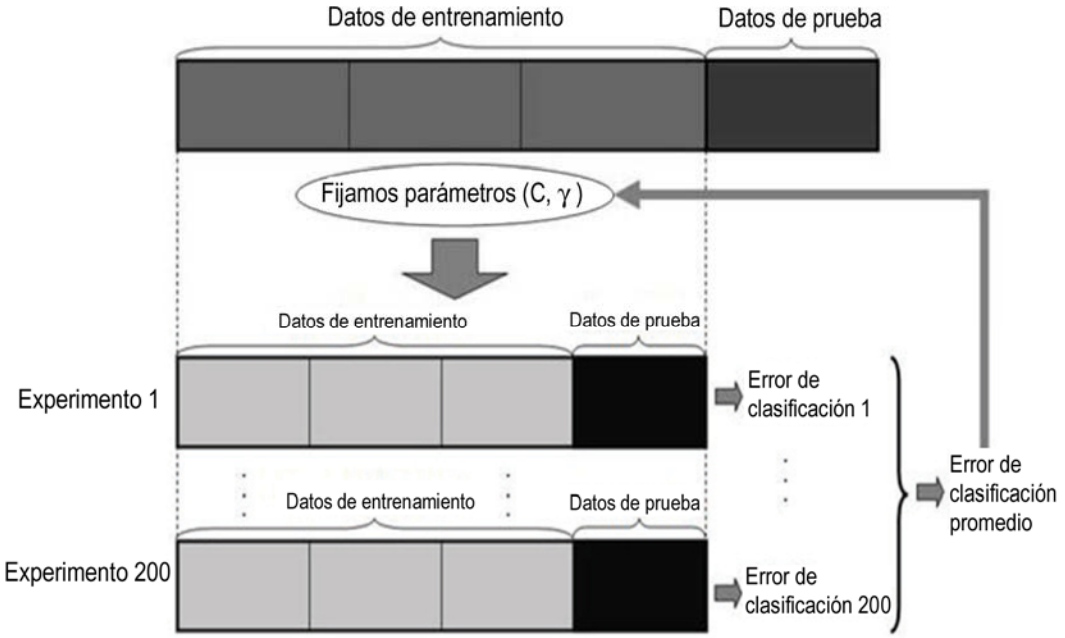

Figura 4. Esquema para la estimación de los parámetros.

Figure 4. Diagram for parameters estimation.

Tabla 1. Estadística descriptiva de los descriptores acústicos por especie (414 anchovetas y 442 sardinas común): mínimo (Xmin), máximo (Xmax), media ($\bar{X}$), desviación estándar ($S(\chi)$) y coeficiente de variación (CV(x)).

Table 1. Statistics minima (Xmin), maxima (Xmax), mean ($\bar{X}$), standard deviation ($S(\chi)$) and coefficient of variation (CV(x)) by species (414 anchovy and 442 common sardine).

Descriptores acústicos	Xmin		Xmax		$\bar{X}$		$S(\chi)$		CV(X)	
	Anchoveta	Sardina común	Anchoveta	Sardina común	Anchoveta	Sardina común	Anchoveta	Sardina común	Anchoveta	Sardina común
Altura media (m)	1,2	1,2	15,2	14,8	3,61	3,41	2,17	2,1	0,6	0,62
Largo (m)	3,58	1,65	2.474,08	1.368,23	38,53	47,74	153,11	131,23	3,97	2,75
Perímetro (m)	10,11	7,19	4.957,61	3.646,83	91,73	127,75	308,19	357,56	3,36	2,8
Área (m^2)	2,82	2,5	12.161,18	13.196,16	153,85	275,03	839,85	1.316,64	5,46	4,79
Elongación	1,1	0,8	434,5	148	10,6	12,05	26,4	18,36	2,49	1,52
Dimensión fractal	1,01	1,02	1,79	1,74	1,3	1,34	0,13	0,15	0,1	0,11
Profundidad de fondo (m)	23	35	703	328	111,48	71,98	95,83	31,52	0,86	0,44
Profundidad cardumen (m)	7,9	7,8	54,9	60,6	12,29	22,51	6,92	19,44	0,56	0,86
Índice altitud	2,35	1,44	98,07	96,55	79,23	62,89	18,78	35,46	0,24	0,56
Energía acústica (m^2 mn^{-2})	9,53	10,64	22.291,97	371.894,70	1.126,41	4.989,83	2.306,55	22.461,88	2,05	4,5
Densidad acústica (1 mn^{-2})	0,01	0,02	163,45	265,87	11,59	20,06	13,74	22,62	1,19	1,13
Distancia costa (mn)	1,04	0,83	17,17	20,41	7,05	3,94	5,19	3,48	0,74	0,88

En la Tabla 2 se presentan los resultados de los errores de clasificación para las etapas de entrenamiento y comprobación a partir de distintas combinaciones de parámetros. En la amplia gama inicial de combinaciones iteradas se escogió un conjunto de nueve posibles combinaciones de parámetros, estas fueron seleccionadas dentro del mejor conjunto de combinaciones encontradas. El error promedio de clasificación más bajo, correspondiente a la etapa de entrenamiento del modelo, fue de 7,05%, con los valores de C = 450 y γ = 0.95. Este par de parámetros asegura además, el menor error en la etapa de comprobación igual a 4,7%. Este último resultado definió los valores de los dos parámetros de la SVM escogidos para clasificar las especies. Los errores de comprobación final entre el conjunto de combinaciones seleccionadas fluctuaron entre 4,7 y 6,5%.

En la Tabla 3 se muestran las matrices de confusión que permite evaluar las incidencia de cada una de las combinaciones de parámetros C y γ del proceso de calibración en la clasificación de las especies. Se apreció que la diferencia en la tasa de aciertos entre los pares de combinaciones varió entre 10 a 14 ejemplos mal clasificados. Las tasas de clasificación más altas tanto para anchoveta como sardina común fueron de 95,3%, respectivamente, cuando los parámetros estimados fueron C = 450 y γ = 0,95.

La Figura 5 representa los resultados del análisis de sensibilidad, en que se muestra que el descriptor batimétrico profundidad del fondo fue el más sensible al SVM debido a que el cociente entre error del modelo al omitir este descriptor y el error del modelo con la totalidad de las variables fue el de mayor valor, igual a 1,88. La segunda variable en importancia fue distancia a la costa, con un valor del cociente igual a 1,798 y la tercera en importancia fue el índice de altitud con 1,259. Por otro lado, se observaron descriptores cuya importancia fue baja al omitirlo del modelo de clasificación. Este es el caso de los descriptores perímetro y altura media, de los cardúmenes cuyos valores fueron de 1,002 y 1,005 respectivamente. Los descriptores morfométricos en general no presentaron una incidencia importante en la clasificación. En relación a los tres indicadores más importantes deducidos del análisis de sensibilidad, la Tabla 1 mostró que para cada especie la profundidad de fondo, distancia a la costa y el índice de altitud estuvieron entre los descriptores con menor CV y donde una de las especies, en este caso la anchoveta presentó un valor promedio menor que la sardina común. En relación a los descriptores con menor importancia (morfométricos), dos de ellos, altura media y dimensión fractal, presentaron bajos CV al igual que los descriptores de mayor importancia, sin embargo, las especies mostraron promedios similares.

DISCUSION

El propósito de este estudio fue efectuar la clasificación acústica de la anchoveta (*Engraulis ringens*) y sardina común (*Strangomera bentincki*) con el método SVM y medir el efecto en la clasificación de las especies debido a la calibración de dos parámetros: el correspondiente al Kernel-Gaussiano, γ y el parámetro de penalización del modelo, C.

Tabla 2. Errores para distintos parámetros en SVM: Kernel-Gaussiano, γ y de penalización, C.
Table 2. Errors for different parameters in SVM: Gaussian-Kernel, γ and Penalty, C.

Parámetros		Error de entrenamiento			Error
C	γ	Mínimo	Máximo	Promedio	Comprobación Final
440	0,94	0,031056	0,130435	0,075714	0,060748
450	0,94	0,037267	0,136646	0,076459	0,060748
460	0,94	0,018634	0,142857	0,072049	0,065421
440	0,95	0,031053	0,124224	0,077142	0,056075
450	0,95	0,024845	0,130435	0,070497	0,046729
460	0,95	0,031056	0,149068	0,073229	0,056075
440	0,96	0,024845	0,161491	0,072608	0,056075
450	0,96	0,031056	0,118012	0,073602	0,065421
460	0,96	0,024845	0,136646	0,072111	0,051402

Tabla 3. Matrices de confusión para un conjunto de parámetros, C y γ .
Table 3. Confusion matrix for a set of parameters, C and γ.

C	γ	Observado	Sardina	Anchoveta	Pronosticado total	Porcentaje de acierto (%)
440	0,94	Sardina	103	4	107	96,4
		Anchoveta	9	98	107	91,6
					214	93,9
440	0,95	Sardina	103	4	107	96,3
		Anchoveta	8	99	107	92,5
					214	94,4
440	0,96	Sardina	102	5	107	95,3
		Anchoveta	7	100	107	93,5
					214	94,4
450	0,94	Sardina	104	3	107	97,2
		Anchoveta	10	97	107	90,7
					214	93,9
450	0,95	Sardina	102	5	107	95,3
		Anchoveta	5	102	107	95,3
					214	95,3
450	0,96	Sardina	100	7	107	93,5
		Anchoveta	7	100	107	93,5
					214	93,5
460	0,94	Sardina	100	7	107	93,5
		Anchoveta	7	100	107	93,5
					214	93,5
460	0,95	Sardina	102	5	107	95,3
		Anchoveta	7	100	107	93,5
					214	94,4
460	0,96	Sardina	102	5	107	95,3
		Anchoveta	6	101	107	94,4
					214	94,9

En todas las matrices de confusión, la sardina fue la especie mejor clasificada, siendo clasificada erróneamente 45 veces de 963 posibilidades en el conjunto de los experimentos, independiente del par de parámetros usados. Los porcentajes de clasificación correcta fluctuaron entre 93,2 y 95,3%. El mejor modelo con C = 450 y γ = 0,95 dio un porcentaje de acierto cercano al 95,3%, clasificando incorrectamente a 10 de 214 cardúmenes. Se puede concluir que el protocolo experimental de calibración mostró una convergencia en los porcentajes de acierto en torno a las diferentes combinaciones de parámetros obtenidos, lo mismo ocurrió con relación a la presencia de la anchoveta como la más frecuentemente mal clasificada. La mejor elección del par de parámetros C y γ dependió del menor error promedio de comprobación. Debido a que el modelo SVM es sensible a cambios en los parámetros tanto del Kernel γ como del de penalización C, se debe poner especial esfuerzo para realizar la mejor elección. Se debe también considerar que la calibración de estos parámetros incide en un costo adicional por mayor consumo de tiempo computacional, especialmente, si la clasificación se basa en un problema multiclase.

Robotham *et al*. (2010) usando un modelo multiclase SVM con tres especies donde se incluye a la especie jurel (*Trachurus murphyi*) obtuvieron porcentajes de clasificación entre 88,8 y 89,3% para sardina común y entre 88,8 y 90,8% para anchoveta, dependiendo del tipo de estrategia de clasificación utilizada (una especie contra otra especie (1-vs-1) y una especie contra el resto de las especies (1-vs-R)). La discriminación obtenida con el modelo SVM basado en dos especies (95,3% de clasificación correcta) fue más efectiva que las obtenidas con el método multiclase

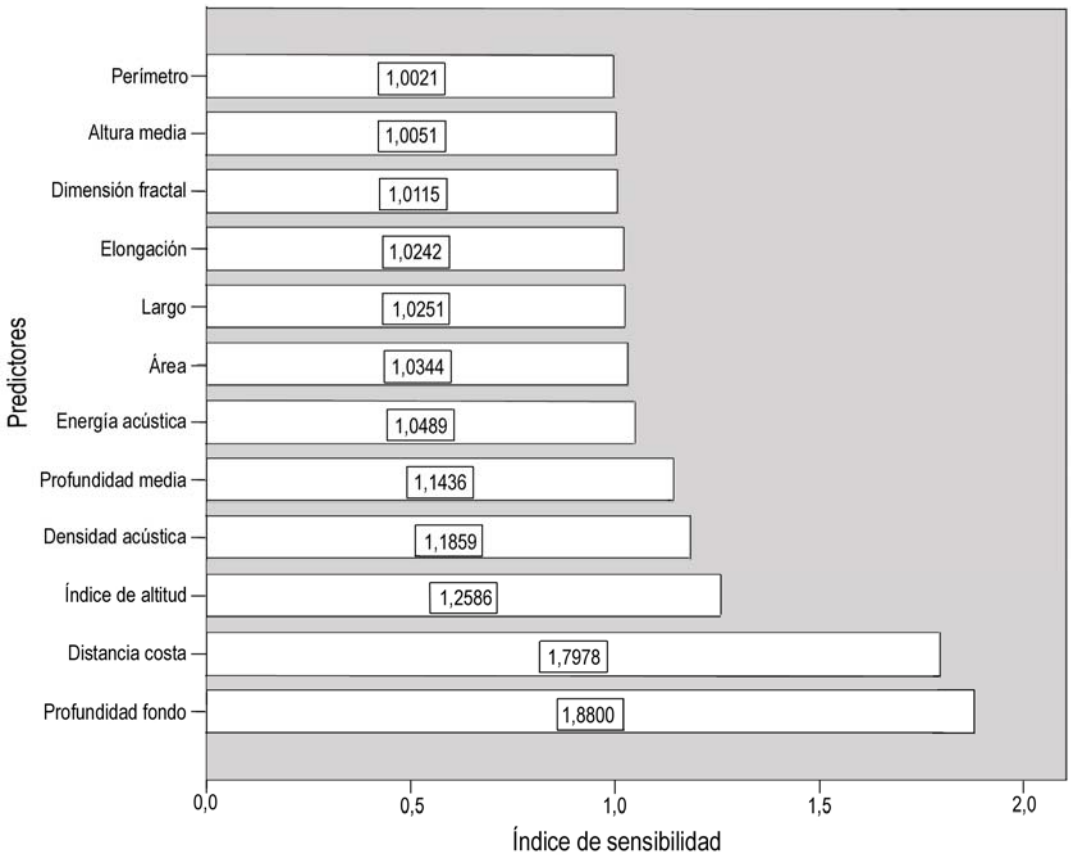

Figura 5. Índice de sensibilidad obtenida como cociente entre error del modelo al omitir un descriptor a la vez y el error del modelo con la totalidad de las variables.

Figure 5. Ratio values **(**sensitivity index) between the classification error when one descriptor at time is removed and the classification error with all descriptors included.

SVM. La menor exactitud indicada se debe a que el método multiclase utilizó en el proceso de clasificación una tercera especie (jurel), lo que genera funciones de decisión más complejas, adicionalmente la muestra utilizada en el proceso de calibración se encontraba desbalanceada, factores que sin duda podrían incidir en la menor clasificación. Estudios similares sobre estas especies en el área no han sido reportados, sin embargo, otros estudios con otros métodos y especies reportan tasas de clasificación entre 77 y 96% (Haralabous & Georgakarakos, 1996; Simmonds *et al*., 1996; Lawson *et al*., 2001; Cabreira *et al*., 2009; Korneliussen *et al*., 2009; Fernandes, 2009).

Los SVM al igual que las redes neuronales se comportan como una caja negra, no es posible observar directamente como incide cada descriptor del cardumen en la clasificación, para lo cual se requiere estimular el análisis y medir el error de clasificación, en este caso mediante la eliminación de un descriptor por vez, y comparar así el efecto de sensibilidad de la variable descriptora en la clasificación. Los resultados del análisis de sensibilidad muestran que el descriptor batimétrico profundidad del fondo es el más sensible al SVM, ya que entrega la mayor contribución al porcentaje de error. La segunda variable en importancia es distancia a la costa.

También resulta interesante hacer notar aquellos descriptores que menos influyen en el error de clasificación una vez que estos han sido eliminados; por ejemplo, es el caso del perímetro y la altura media del cardumen, así como el resto de los descriptores morfológicos, tiene como explicación que la anchoveta y la sardina forman cardúmenes morfológicamente similares en cuanto a estos descriptores. La anchoveta y la sardina común tienen similares propiedades acústicas y características biológicas, ambas especies co-habitan la misma área ecológica (cercana a la costa) con sólo escasas diferencias entre ellas. En este estudio los descriptores morfológicos tienen baja incidencia en la clasificación siendo un resultado consistente con los resultados de Korneliusse *et al*. (2009), que clasifican datos

acústicos con multifrecuencia apoyada en información morfológica y distribución geográfica de las especies. Los análisis estadísticos de los descriptores, en base principalmente al coeficiente de variación y la media de las especies, proporcionan evidencias numéricas consistentes con la identificación de las variables consideradas efectivas y no efectivas para clasificar que fueron detectadas a partir del análisis de sensibilidad. En general los descriptores con CV bajos y diferencias de medias importantes entre especies coinciden con las variables más efectivas para la clasificación. Por otra parte, descriptores con CV bajos y con medias similares entre especies no mostraron ser variables efectivas importantes. Descriptores con CV altos y medias diferentes entre especies, mostraron también no ser tan efectivas para clasificar. El comportamiento preliminar observado de la relación CV y media de los descriptores indican un comportamiento consistente con el análisis de sensibilidad, que requiere de una mayor cantidad de experimentos para obtener una relación más precisa y concluyente.

Aprovechando la potencialidad del método (SVM), se puede utilizar para resolver un problema de clasificación de especies marinas evaluadas acústicamente. Es importante notar que la estructura de la SVM a menudo implica, la optimización cuadrática convexa, lo que lleva a tener soluciones globales y únicas. Es un método no paramétrico que no requiere supuestos estadísticos sobre las variables o predictores. El método SVM ajustó correctamente el 95,3% de los cardúmenes de anchoveta y sardina. La especie con menor error de clasificación fue la sardina común. Los parámetros del Kernel-Gaussiano γ y de penalización C van a incidir en la matriz de confusión y los porcentajes de clasificación final, por lo que se sugiere establecer un protocolo experimental de calibración que permita realizar de manera cuidadosa de su elección, teniendo en consideración los mayores costos por tiempo de procesos. Este último aspecto se hace más crítico cuando la aplicación es multiclase. El esfuerzo necesario (tiempo de proceso) para escoger y optimizar la función Kernel para producir la clasificación más eficiente es la mayor dificultad del método SVM.

En general, el método SVM trabaja bien con aplicaciones en diferentes ámbitos como es el campo de la identificación de especies en ecología (Morris *et al*., 2001) y otras aplicaciones biológicas como clasificación de automática de edad de los peces usando imágenes de otolitos (Bermejo *et al*., 2007).

AGRADECIMIENTOS

Los autores agradecen los comentarios e importantes sugerencias realizadas al trabajo por dos revisores. Este trabajo utilizó datos obtenidos por el Instituto de Fomento Pesquero de proyectos financiados por la Subsecretaría de Pesca de Chile.

REFERENCIAS

Barange, M. 1994. Acoustic identification, classification and structure of biological patchiness on the edge of the Agulhas Bank and its relationship to frontal features. S. Afr. J. Mar. Sci., 14: 333-347.

Bermejo, S., B. Monegal & J. Cabestany. 2007. Fish age categorization from otolith images using multi-class support vector machines. Fish. Res., 84: 247-253.

Boser, B.E., I.M. Guyon & V. Vapnik. 1992. A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on computational learning theory. Pittsburg, ACM, pp. 144-152.

Buelens, B., T. Pauly, R. Williams & A. Sale. 2009. Kernel methods for the detection and classification of fish schools in single-beam and multibeam acoustic data. ICES J. Mar. Sci., 66: 1130-1135.

Cabreira, A.G., M. Tripode & A. Madirolas. 2009. Artificial neural networks for fish-species identification. ICES J. Mar. Sci., 66: 1119-1129.

Coetzee, J. 2000. Use of a shoal analysis and patch estimation system (SHAPES) to characterize sardine schools. Aquat. Living Resour., 13: 1-10.

Demer, D.A., G.R. Cutter, J.S. Renfree & J.L. Butler. 2009. A statistical-spectral method for echo classification. ICES J. Mar. Sci., 66: 1081-1090.

Fablet, R., R. Lefort, I. Karoui, L. Merger, J. Massé, C. Scalabrin & J. M Boucher. 2009. Classifying fish schools and estimating their species proportions in fishery-acoustics surveys. ICES J. Mar. Sci., 66: 1136-1142.

Fernandes, P.G., R.J. Korneliussen, A. Lebourges-Dhaussy, J. Masse, M. Iglesias, N. Diner & E. Ona. 2006. The SIMFAMI proyect: species identification methods from acoustics multifrecuency information. Final Report to the EC, Number Q5RS-2001-02054.

Fernandes, P.G. 2009. Classification trees for species identification of fish-school echotraces. ICES J. Mar. Sci., 66: 1073-1080.

Foote, K., H. Knudsen, G. Vestnes, D. MacLennan & J. Simmonds. 1987. Calibration of acoustics instruments for fish density estimation: a practical guide. ICES Coop. Res. Rep., 144: 57 pp.

Fleischman, S.J. & D.L. Burwen. 2003. Mixture models for the species apportionment of hidroacoustics data,

with echo-envelope length as the discriminatory variable. ICES J. Mar. Sci., 60: 592-598.

Hastie, T., R. Tibshirirani & J. Friedman. 2001. The elements of statistical learning. Springer-Verlag, New York, 533 pp.

Haralabous, J. & S. Georgakarakos. 1996. Artificial neural networks as a tool for species identification of fish schools. ICES J. Mar. Sci., 53: 173-180.

Horne, J.K. 2000. Acoustic approaches to remote species identification: a review. Fish. Oceonogr., 94: 356-371.

Hsu, C.W. & C.J. Lin. 2002. A comparison of methods for multi-class support vector machines. IEEE Trans. Neural Net., 13: 415-425.

Joachims, T. 2001. Learning to classify text using support vector machines. Methods, theory and algorithms. Kluwer Academic Publishers, London, 205 pp.

Korneliussen, R.J., Y. Heggelund, I.K. Eliassen & G.O. Johansen. 2009. Acoustics species identification of schooling fish. ICES J. Mar. Sci., 66: 1111-1118.

Lawson, G.L., M. Barange & P. Fréon. 2001. Species identifiation of pelagic fish schools on the South African continental shelf using acoustic descriptors and ancillary information. ICES J. Mar. Sci., 58: 275-287.

Morris, C.W., A. Audret & I. Boddy. 2001. Support vector machines for identifying organisms: a comparison with strongly partitioned radial basis function networks. Ecol. Model., 146: 57-67.

Nero, R.W. & J.J. Magnuson. 1989. Characterization of patches along transects using high-resolution 70-kHz integrated acoustics data. Can. J. Fish. Aquat. Sci., 46: 2056-2064.

Richards, L.J., R. Kieser, T.J. Mulligan & J.R. Candy. 1991. Classification of fish assemblages based on echo integration surveys. Can. J. Fish. Aquat. Sci., 48: 1264-1272.

Robotham, H., P. Bosch, J.C. Gutiérrez-Estrada, J. Castillo & I. Pulido-Calvo. 2010. Acoustics identification of small pelagic fish species in Chile using support vector machines and neural networks. Fish. Res., 102: 115-122.

Received: 23 December 2010; Accepted: 2 January 2012

Scalabrin, C., 1991. Recherche d'une méthodologie pour la classification et l´identification automatiques des détections acoustics des bancs de poissons. Rapp. IFREMER, DITI/NPA 91.23.

Scalabrin, C. & J. Massé. 1993. Acoustic detection of the spatial and temporal distribution of fish shoals in the Bay of Biscay. Aquat. Living Resour., 6: 269-283.

Scalabrin, C., N. Diner, A. Weill, A. Hillion & M.C. Mouchot. 1996. Narrowband acoustic identification of monospecific fish shoals. ICES J. Mar. Sci., 53: 181-188.

Scholkopf, B. & A. Smola. 2002. Learning with kernels. The MIT Press, Cambridge, 626 pp.

Simmonds, E.J. & D.N. MacLennan, 2005. Fisheries acoustics: theory and practice. Fish and Aquatic Resources Series 10. Blackwell Science, Oxford, 437 pp.

Simmonds, E.J., F. Armstrong & P.J. Copland. 1996. Species identification using wideband backscatter with neural network and discriminant analysis. ICES J. Mar. Sci., 53: 189-195.

Tegowski, J., N. Gorska & Z. Klusek. 2003. Statistical analysis of acoustic echoes from underwater meadows in the eutrophic Puck Bay (southern Baltic Sea). Aquat. Living Resour., 16: 215-221.

Vapnik, V.N. 1995. The nature of statistical learning theory. Springer, New York, 315 pp.

Vray, D., G. Gimenez & R. Person. 1990. Attempt at classification of echo-sounder signals based on the linear discrimination function of Fisher. Rapp. P.-V. Réun. Cons. Int. Explor. Mer, 189: 388-393.

Weston, J. & C. Watkins. 1998. Multi-class support vector machines. Department of computer Science, royal Holloway, University of London, Technical Report CSD-TR-98-04, 9 pp.