Capítulo 5

Máquina de Aprendizaje *K*-SVCR

"Cualquier idea simple se puede expresar de la manera más compleja que existe"
Ley de Murphy.

Para comenzar

Las máquinas de soporte vectorial que trabajan en problemas de clasificación están definidas de forma específica para tratar problemas de clasificación binaria. Un nuevo algoritmo basado en vectores soporte será definido con el objetivo posterior de ser utilizado durante el esquema de descomposición de multiclasificación, denominado algoritmo K-SVCR. Cuando se trata un problema de clasificación multiclase, $K > 2$, toma sentido construir máquinas de aprendizaje que asignen salida $+1$ o -1 si el patrón de entrenamiento pertenece a las clases a ser separadas, y salida 0 si por contra el patrón tiene una etiqueta diferente a las anteriores. Así se está forzando al hiperplano separador de las dos clases implicadas no sólo a poseer características de margen máximo, sino también estar restringido a recubrir todos los patrones de entrenamiento "0-etiquetados", aquellos pertenecientes a cualquier otra clase que no sean las dos iniciales. El problema de programación cuadrática restringido asociado podría ser interpretado como un intermedio entre el método SVMC y el método SVMR de entrenamiento sobre vectores soporte para estimación de regresiones.

5.1 Motivación e Interpretación

El análisis realizado en el Capítulo anterior sobre la definición, funcionamiento y características que poseen diferentes arquitecturas de clasificación multiclase basadas

en clasificadores binarios, en concreto SVMCs, permite elaborar un esquema de propiedades que sería deseable que cumpliera una máquina de multiclasificación ideal. Estas propiedades podrían resumirse en:

1. Un esquema de descomposición de baja complejidad, que se traduciría en el uso de un número reducido de dicotomías.
2. Un problema de optimización asociado de baja dimensión, que permitiría obtener resultados con un coste computacional no muy elevado.
3. Un conjunto de entrenamiento para cada dicotomía equilibrado y de tamaño similar al conjunto de aprendizaje original.
4. Robustez de las respuestas ante fallos parciales de los nodos de dicotomía.
5. Una base teórica que permita asegurar una cota baja en el error de generalización.
6. Un tiempo de evaluación de resultados lo más reducido posible.

Muchas de estas propiedades son opuestas las unas de las otras. Un esquema tipo "todas las clases a la vez" posee una descomposición sencilla, una única máquina, pero el problema de optimización cuadrática es de gran dimensionalidad. Una mejora de robustez del método 1-v -1 mediante técnicas de ECOC implica un aumento significativo en el número de dicotomías del esquema de descomposición. De igual forma que el entrenamiento de una dicotomía sobre todo el conjunto de aprendizaje implica un coste computacional más elevado, el objetivo de basar la clasificación multiclase en árboles que posean una base teórica de buena generalización y reducción en el tiempo de evaluación provocan una arquitectura nada robusta.

Hallar la mejora de algunas de estas características de comportamiento, sin provocar una reducción en el grado de cumplimiento de las restantes es el objetivo que motiva la búsqueda de una estructura equilibrada respecto a todas las propiedades, de la que pueda preverse un mejor comportamiento global, sin necesidad de hacer referencia explícita a un tipo u otro de 'benchmark de validación', aunque su comportamiento sobre estos problemas deberá ser en general mejor que el resto de arquitecturas.

5.1.1 Idea Motivadora

Una vez analizado en profundidad el problema del aprendizaje multiclase desde la perspectiva de buena generalización que ofrecen los biclasificadores SVMC, se ha elegido como esquema base de descomposición el método 1-v-1 debido a que las dicotomías obtenidas con esta formulación permiten concentrar el esfuerzo clasificador

sobre una zona determinada del espacio de clasificación. No hay que olvidar sin embargo que la exclusión de la información aportada por las entradas pertenecientes al resto de clases provoca que la fisonomía del hiperplano de decisión no se adapte a éstas.

La idea que permitirá incluir estas otras clases en el entrenamiento de la dicotomía es una generalización de una propiedad muy empleada en el cálculo de SVMCs. Supóngase que un usuario intenta resolver el problema cuadrático de optimización primal asociado a una máquina SVMC, de función objetivo (3.6), con restricciones (3.4)–(3.5) que conduce a una función de decisión (3.3) basada en un hiperplano óptimo del tipo (3.1). Tal como se afirma por ejemplo en [Burges, 1998], es muy común suponer que el término independiente en (3.1) es nulo, $b = 0$, por lo que el hiperplano óptimo hallado será del estilo

$$f(\mathbf{x}, \omega) = \langle \omega, \mathbf{x} \rangle_{\mathcal{F}} = k(\omega, \mathbf{x}) \,. \tag{5.1}$$

Con esta suposición se consigue que la restricción de igualdad (3.9) obtenida al buscar el punto de silla del lagrangiano respecto al término independiente,

$$\frac{\partial W(\alpha)}{\partial b} = 0 \,, \tag{5.2}$$

no aparezca entre las restricciones del enunciado dual (3.7). La finalidad última es puramente operativa, se obtiene una expresión de las restricciones a (3.7) en forma de desigualdades sencillas y no es necesario realizar una restricción igualdad costosa como (3.9) que es un sumatorio de productos sobre todas las salidas de $\mathcal{T}$, ℓ. El coste de esta suposición es mínimo, geométricamente hablando significa obligar al hiperplano óptimo a que pase por el origen de coordenadas en el espacio de características $\mathcal{F}$. Puesto que este espacio siempre es elegido para que su dimensión sea mucho mayor que la del original a fin de convertir en separables conjuntos que en el espacio original no lo eran,

$$\dim \mathcal{F} \gg d = \dim \mathbb{R}^d \,, \tag{5.3}$$

la suposición $b = 0$ es inapreciable en el resultado final pues sólo supone la reducción del espacio de características en 1 dimensión.

Siguiendo este procedimiento, pensado inicialmente para facilitar el tratamiento numérico, si el espacio de características es lo bastante amplio, no sólo 1, muchas restricciones de este estilo afectarán de forma poco apreciable el resultado final. La propuesta por tanto podría expresarse en los siguientes términos.

Definición 5.1. *Dado el conjunto de entrenamiento $\mathcal{T}$ definido en (4.1) se define como* función de decisión ternaria $h(\mathbf{x})$*, aquella basada en un hiperplano de decisión $f(\mathbf{x}, \omega)$ cumpliendo:*

$$h(\mathbf{x}_p) = \begin{cases} +1 & p = 1, \ldots, \ell_1 \\ -1 & p = \ell_1 + 1, \ldots, \ell_1 + \ell_2 \\ 0 & p = \ell_1 + \ell_2 + 1, \ldots, \ell \end{cases} \,, \tag{5.4}$$

donde, sin pérdida de generalidad, se han supuesto los primeros ℓ_1 y ℓ_2 patrones ($\ell_{12} = \ell_1 + \ell_2$) correspondiendo a las dos clases a ser separadas, mientras que el resto de patrones ($\ell_3 = \ell - \ell_{12}$) pertenecen a cualquier otra clase diferente — serán etiquetadas con 0 *—.*

Obviamente, no existe, en general, ningún hiperplano en el espacio de entrada $\mathcal{X} \subseteq \mathbb{R}^d$ que permita definir una función de decisión ternaria (5.4) de tales características, y por tanto es inútil buscar una solución lineal al problema en este espacio. Sin embargo, al ser insertado este espacio original por medio de una inclusión no lineal dentro de un espacio de características de dimensión suficientemente grande, la capacidad de los hiperplanos para generar una función de decisión que cumpla las condiciones de la definición se incrementa y hará posible hallar una solución. El requerimiento pretendido por la máquina de aprendizaje que se quiere definir es obligar al hiperplano óptimo a contener todos los ℓ_3 patrones de entrenamiento con etiqueta 0.

5.2 Máquina de Aprendizaje *K*-SVCR

La definición de la función de decisión ternaria sobre un planteamiento de problema de clasificación con tres clases crea la necesidad de definir una máquina de aprendizaje capaz de triclasificar basada en las SVMCs. Siguiendo el método de exposición usual, se comenzará por definir la nueva máquina en el caso separable para luego generalizar al caso no separable. De hecho, la exposición de máquinas SVMC suele comenzar con el caso linealmente separable y continuar con el no linealmente no separable. En la presente formulación se hace siempre necesario introducir el espacio de entrada $\mathcal{X}$ en un espacio de características de mayor dimensión para asegurar el éxito de la clasificación por lo que los casos lineales no son tratados[1] .

5.2.1 Caso No Lineal Separable

Definición 5.2. *Se define el problema de optimización restringida asociado a la máquina de clasificación triclase K-SVCR, para el caso separable como: para* $0 \leq \delta < 1$ *elegido a priori,*

$$\arg\min R_{sep}(\omega) = \frac{1}{2} \|\omega\|^2_{\mathcal{F}} \,, \tag{5.5}$$

sujeto a

$$y_i \cdot (\langle \omega, \mathbf{x}_i \rangle_{\mathcal{F}} + b) - 1 \geq 0 \qquad i = 1, \ldots, \ell_{12} \,, \tag{5.6}$$

$$-\delta \leq \langle \omega, \mathbf{x}_i \rangle_{\mathcal{F}} + b \leq \delta \qquad i = \ell_{12} + 1, \ldots, \ell \,, \tag{5.7}$$

[1] En caso que un separador lineal bastase, el separador no lineal construido por la nueva máquina adoptará una fisonomía quasilineal en la zona cercana a los patrones de entrenamiento.

que permite definir la K-SVCR función de decisión solución similar a (5.4), definida por

$$h(\mathbf{x}) = \begin{cases} +1 & si \quad f(\mathbf{x},\omega) = \langle \omega, \mathbf{x} \rangle_{\mathcal{F}} + b > \delta \\ -1 & si \quad f(\mathbf{x},\omega) = \langle \omega, \mathbf{x} \rangle_{\mathcal{F}} + b < -\delta \\ 0 & en\ caso\ contrario \end{cases} . \tag{5.8}$$

La nueva metodología es consistente con la máquina estándar de biclasificación SVMC,

Proposición 5.3. *Si $K = 2$, lo que implica $\ell_3 = 0$, entonces los problemas de optimización SVMC y K-SVCR son equivalentes.*

Proposición 5.4. *Siendo $K > 2$, si se cumple que $\delta = 0$, entonces la función decisión (5.8) es la misma que la propuesta originalmente (3.3) para las máquinas SVMC.*

Si en la Ecuación 5.8 de la definición anterior se hubiese propuesto una salida no nula admitiendo desigualdades no estrictas y se utilizase un valor $\delta = 0$ como en la Proposición 5.4, se estaría requiriendo que el hiperplano separador contenga los últimos ℓ_3 patrones de entrenamiento de forma exacta. Sin embargo, esta imposición implica:

- La no generalización de las respuestas para la clase etiquetada 0.
- El número de vectores soporte será muy elevado entre los patrones con etiqueta 0.[Smola, 1998]
- Mayor coste computacional.

Al elegir una $\delta > 0$, aunque nuestra tarea es el reconocimiento de patrones, podría decirse que se ha realizado un cierto uso de la función de coste ε-insensitiva (3.12) de la página 54 empleada en el método SVMR, sobre salidas de tipo $y_i = 0$. Siguiendo esta línea, la función de decisión ternaria (5.8) podría reescribirse como

$$h(\mathbf{x}) = sign\left(f(\mathbf{x},\omega) \cdot |f(\mathbf{x},\omega)|_{\delta}\right) . \tag{5.9}$$

Debe hacerse hincapié en el hecho de que si bien el nuevo método podría definirse como una 1-v-1 SVMC con ayuda de metodología SVMR para la clasificación ternaria, de ninguna manera se trata de una SVMR con tres salidas con fines de clasificación. En tal caso, el primer grupo de restricciones (5.6) sería substituido por este otro

$$-\delta \leq y_i \cdot (\langle \omega, \mathbf{x}_i \rangle_{\mathcal{F}} + b) - 1 \leq \delta \quad i = 1, \ldots, \ell_{12} , \tag{5.10}$$

lo que supondría un aumento significativo en el número de restricciones y un espacio de convexidad donde optimizar el vector de parámetros mucho menor ya que las

nuevas restricciones son mucho más exigentes que las anteriores. No sólo los patrones 0-etiquetados deberían estar situados en una banda o tubo de diámetro 2δ en el espacio de características, sino que también los patrones con etiquetas ± 1 deberían estar situados en tubos semejantes centrados en $+1$ y -1, paralelos al tubo centrado en 0 en el espacio de características. Resulta evidente que esta formulación con base SVMR es mucho más exigente que la máquina K-SVCR propuesta y producirá soluciones con menor poder generalizador.

Para un estudio sobre condiciones que permiten derivar una SVMC como un caso especial de SVMR puede consultarse [Pontil et al., 1998].

Obtención de la Solución

Siguiendo la demostración realizada en [Angulo and Català, 2000a], la solución para el problema definido en (5.5) con restricciones (5.6)–(5.7) puede ser hallada localizando el punto de silla del lagrangiano

$$\begin{aligned} L\left(\omega, b, \alpha, \beta, \beta^*\right) = \frac{1}{2}\left\|\omega\right\|_{\mathcal{F}}^2 - \sum_{i=1}^{\ell_{12}} \alpha_i \left[y_i \left(\left\langle \omega, \mathbf{x}_i \right\rangle_{\mathcal{F}} + b\right) - 1\right] + \\ + \sum_{i=\ell_{12}+1}^{\ell} \beta_i \left[y_i \left(\left\langle \omega, \mathbf{x}_i \right\rangle_{\mathcal{F}} + b\right) - \delta\right] - \\ - \sum_{i=\ell_{12}+1}^{\ell} \beta_i^* \left[y_i \left(\left\langle \omega, \mathbf{x}_i \right\rangle_{\mathcal{F}} + b\right) - \delta\right] \end{aligned} , \tag{5.11}$$

con restricciones

$$\begin{aligned} \alpha_i \geq 0 \quad i = 1, \ldots, \ell_{12} \\ \beta_i, \beta_i^* \geq 0 \quad i = \ell_{12}+1, \ldots, \ell \end{aligned} , \tag{5.12}$$

que debe ser maximizado respecto a las variables duales α_i y β_i, β_i^*, y minimizado respecto a las variables primales ω y b. En el punto de silla la solución debería satisfacer las condiciones:

$$\begin{aligned} \frac{\partial}{\partial \omega} L\left(\omega, b, \alpha, \beta, \beta^*\right) = 0 \\ \frac{\partial}{\partial b} L\left(\omega, b, \alpha, \beta, \beta^*\right) = 0 \end{aligned} , \tag{5.13}$$

que conducen a las ecuaciones

$$\begin{aligned} \omega = \sum_{i=1}^{\ell_{12}} \alpha_i y_i \mathbf{x}_i - \sum_{i=\ell_{12}+1}^{\ell} \left(\beta_i - \beta_i^*\right) \mathbf{x}_i \\ 0 = \sum_{i=1}^{\ell_{12}} \alpha_i y_i - \sum_{i=\ell_{12}+1}^{\ell} \left(\beta_i - \beta_i^*\right) \mathbf{x}_i \end{aligned} . \tag{5.14}$$

Finalmente, si se definen las variables de ayuda

$$\gamma_i = \begin{cases} \alpha_i y_i & i = 1, \ldots, \ell_{12} \\ \beta_i & i = \ell_{12}+1, \ldots, \ell \\ \beta^*_{i-\ell_3} & i = \ell+1, \ldots, \ell+\ell_3 \end{cases}, \tag{5.15}$$

se consigue la eliminación de las variables primales y se obtiene el problema de optimización dual de Wolfe: para $0 \leq \delta < 1$ elegido *a priori*,

$$\arg\min W(\gamma) = \frac{1}{2}\gamma^\top \cdot \mathbf{H} \cdot \gamma + \mathbf{c}^\top \cdot \gamma\,, \tag{5.16}$$

con

$$\gamma^\top = (\gamma_1, \ldots, \gamma_\ell, \gamma_{\ell+1}, \ldots, \gamma_{\ell+\ell_3}) \in \mathbb{R}^{\ell_{12}+\ell_3+\ell_3}\,, \tag{5.17}$$

$$\mathbf{c}^\top = \left(\frac{-1}{y_1}, \cdots, \frac{-1}{y_{\ell_{12}}}, \delta, \ldots, \delta\right) \in \mathbb{R}^{\ell_{12}+\ell_3+\ell_3}\,, \tag{5.18}$$

$$\mathbf{H} = \begin{pmatrix} (k(\mathbf{x}_i, \mathbf{x}_j)) & -(k(\mathbf{x}_i, \mathbf{x}_j)) & (k(\mathbf{x}_i, \mathbf{x}_j)) \\ -(k(\mathbf{x}_i, \mathbf{x}_j)) & (k(\mathbf{x}_i, \mathbf{x}_j)) & -(k(\mathbf{x}_i, \mathbf{x}_j)) \\ (k(\mathbf{x}_i, \mathbf{x}_j)) & -(k(\mathbf{x}_i, \mathbf{x}_j)) & (k(\mathbf{x}_i, \mathbf{x}_j)) \end{pmatrix}, \tag{5.19}$$

$$\mathbf{H} = \mathbf{H}^\top \in \mathcal{M}\left(\mathbb{R}^{\ell_{12}+\ell_3+\ell_3}, \mathbb{R}^{\ell_{12}+\ell_3+\ell_3}\right), \tag{5.20}$$

sujeto a

$$\begin{matrix} \gamma_i \cdot y_i \geq 0 & i = 1, \ldots, \ell_{12} \\ \gamma_i \geq 0 & i = \ell_{12}, \ldots, \ell+\ell_3 \end{matrix}, \tag{5.21}$$

$$\sum_{i=1}^{\ell_{12}} \gamma_i = \sum_{i=\ell_{12}+1}^{\ell} \gamma_i - \sum_{i=\ell+1}^{\ell+\ell_3} \gamma_i\,. \tag{5.22}$$

La función de decisión puede escribirse a partir del hiperplano separador solución como

$$h(\mathbf{x}) = \begin{cases} +1 & \text{si} \quad f(\mathbf{x}, \omega) = \sum_{i=1}^{SV} \nu_i k(\mathbf{x}_i, \mathbf{x}) + b > \delta \\ -1 & \text{si} \quad f(\mathbf{x}, \omega) = \sum_{i=1}^{SV} \nu_i k(\mathbf{x}_i, \mathbf{x}) + b < -\delta \\ 0 & \text{en caso contrario} \end{cases}, \tag{5.23}$$

donde

$$\nu_i = \begin{cases} \gamma_i & i = 1, \ldots, \ell_{12} \\ \gamma_{i+\ell_3} - \gamma_i & i = \ell_{12}+1, \ldots, \ell \end{cases}, \tag{5.24}$$

y b es calculada de (5.7) cuando adopta la forma de igualdad sobre los vectores soporte en términos del vector de parámetros γ. También puede observarse que la restricción (5.22) puede ser escrita como

$$\sum_{i=1}^{SV} \nu_i = 0\,. \tag{5.25}$$

5.2.2 Caso No Lineal No Separable

Definición 5.5. *Se define el problema de optimización restringida asociado al método K-SVCR, para el caso general como: para* $0 \leq \delta < 1$ *elegido a priori,*

$$\arg\min R_{K-SVCR}\left(\omega, \xi, \varphi^{(*)}\right) = \frac{1}{2}\left\|\omega\right\|_{\mathcal{F}}^2 + C\sum_{i=1}^{\ell_{12}}\xi_i + D\sum_{i=\ell_{12}+1}^{\ell}\left(\varphi_i + \varphi_i^*\right), \tag{5.26}$$

sujeto a

$$y_i \cdot \left(\left\langle \omega, \mathbf{x}_i\right\rangle_{\mathcal{F}} + b\right) \geq 1 - \xi_i \qquad i = 1, \ldots, \ell_{12}, \tag{5.27}$$

$$-\delta - \varphi_i^* \leq \left\langle \omega, \mathbf{x}_i\right\rangle_{\mathcal{F}} + b \leq \delta + \varphi_i \qquad i = \ell_{12}+1, \ldots, \ell, \tag{5.28}$$

$$\begin{array}{r} \xi_i \geq 0 \quad i = 1, \ldots, \ell_{12} \\ \varphi_i, \varphi_i^* \geq 0 \quad i = \ell_{12}+1, \ldots, \ell \end{array}. \tag{5.29}$$

Una solución al problema definido por (5.26) y restricciones (5.27)—(5.29) puede ser hallada solucionando el problema de optimización dual de Wolfe definido como: para $0 \leq \delta < 1$ elegido *a priori*,

$$\arg\min W\left(\gamma\right) = \frac{1}{2}\gamma^\top \cdot \mathbf{H} \cdot \gamma + \mathbf{c}^\top \cdot \gamma, \tag{5.30}$$

con

$$\gamma^\top = \left(\gamma_1, \ldots, \gamma_\ell, \gamma_{\ell+1}, \ldots, \gamma_{\ell+\ell_3}\right) \in \mathbb{R}^{\ell_{12}+\ell_3+\ell_3}, \tag{5.31}$$

$$\mathbf{c}^\top = \left(\frac{-1}{y_1}, \cdots, \frac{-1}{y_{\ell_{12}}}, \delta, \ldots, \delta\right) \in \mathbb{R}^{\ell_{12}+\ell_3+\ell_3}, \tag{5.32}$$

$$\mathbf{H} = \begin{pmatrix} \left(k\left(\mathbf{x}_i, \mathbf{x}_j\right)\right) & -\left(k\left(\mathbf{x}_i, \mathbf{x}_j\right)\right) & \left(k\left(\mathbf{x}_i, \mathbf{x}_j\right)\right) \\ -\left(k\left(\mathbf{x}_i, \mathbf{x}_j\right)\right) & \left(k\left(\mathbf{x}_i, \mathbf{x}_j\right)\right) & -\left(k\left(\mathbf{x}_i, \mathbf{x}_j\right)\right) \\ \left(k\left(\mathbf{x}_i, \mathbf{x}_j\right)\right) & -\left(k\left(\mathbf{x}_i, \mathbf{x}_j\right)\right) & \left(k\left(\mathbf{x}_i, \mathbf{x}_j\right)\right) \end{pmatrix}, \tag{5.33}$$

$$\mathbf{H} = \mathbf{H}^\top \in \mathcal{M}\left(\mathbb{R}^{\ell_{12}+\ell_3+\ell_3}, \mathbb{R}^{\ell_{12}+\ell_3+\ell_3}\right), \tag{5.34}$$

restringido a

$$\begin{array}{r} 0 \leq \gamma_i \cdot y_i \leq C \quad i = 1, \ldots, \ell_{12} \\ 0 \leq \gamma_i \leq D \quad i = \ell_{12}+1, \ldots, \ell+\ell_3 \end{array}, \tag{5.35}$$

$$\sum_{i=1}^{\ell_{12}}\gamma_i = \sum_{i=\ell_{12}+1}^{\ell}\gamma_i - \sum_{i=\ell+1}^{\ell+\ell_3}\gamma_i. \tag{5.36}$$

La función de decisión puede ser escrita como

$$h\left(\mathbf{x}\right) = sign\left(f\left(\mathbf{x}, \omega\right) \cdot \left|f\left(\mathbf{x}, \omega\right)\right|_\delta\right), \tag{5.37}$$

a partir del hiperplano solución

$$f(\mathbf{x},\omega) = \sum_{i=1}^{SV} \nu_i k(\mathbf{x}_i,\mathbf{x}) + b \quad , \tag{5.38}$$

donde

$$\nu_i = \begin{cases} \gamma_i & i = 1,\ldots,\ell_{12} \\ \gamma_{i+\ell_3} - \gamma_i & i = \ell_{12}+1,\ldots,\ell \end{cases} \quad , \tag{5.39}$$

y b es calculada a partir de (5.28) cuando adopta la forma de igualdad sobre los vectores soporte en términos de los parámetros γ_i's. Puede ser observado que la restricción (5.36) se reescribe como

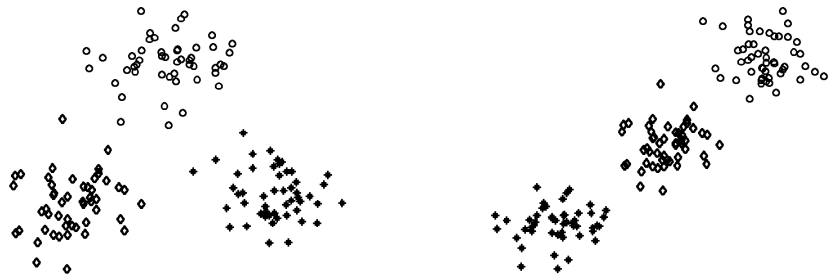
$$\sum_{i=1}^{SV} \nu_i = 0\,. \tag{5.40}$$

El problema general de optimización (5.26) con restricciones (5.27)—(5.29) asociado a la máquina K-SVCR podría ser notado como problema de clasificación Soft-SVMC$_{12}$ + Soft-SVMR$_3$, indicando la fusión entre restricciones del estilo SVMC y SVMR con márgenes débiles sobre tres clases distinguidas de patrones con fines de clasificación. Haciendo nulas las diferentes variables artificiales ξ_i, φ_i, φ_i^*, la solución para los otros dos diferentes problemas de clasificación {Hard/Soft}-SVMC$_{12}$ + {Soft/Hard}-SVMR$_3$ se obtiene de forma directa.

Otras variaciones como situar un nivel δ de ancho de banda diferente para cada punto de entrenamiento [Smola, 1998], o incluso diferente para cada uno de los dos grupos de restricciones en (5.28) también son posibles y llevan a soluciones fáciles de derivar a partir de los dos casos generales aquí presentados.

5.3 Experimentación

Para comprobar las características de funcionamiento y las prestaciones de la nueva máquina de aprendizaje K-SVCR se utilizarán una serie de problemas artificiales tanto para el caso separable como el no separable. El objetivo principal de los experimentos será de una parte analizar el papel desempeñado por el nuevo factor de insensitividad δ introducido en la nueva algorítmica y por otra destacar el mejor comportamiento de la nueva máquina respecto a un diseño de multiclasificación con máquinas SVMR. Se ha de recordar que las máquinas SVMC sólo pueden biclasificar, por lo que la generalización hacia el caso multiclase podría desarrollarse mediante SVMRs. Se probará que la utilización de la máquina SVMR significa llevar la generalización hacia la multiclasificación demasiado allá y que resulta más correcto una máquina intermedia, como la K-SVCR, con restricciones de tipo mixto clasificación y regresión para desarrollar esta tarea.



(a) Conjunto de entrenamiento $\mathcal{T}_1$. (b) Conjunto de entrenamiento $\mathcal{T}_2$.

Figura 5.1: Conjuntos de entrenamiento linealmente separables.

5.3.1 Problemas Artificiales

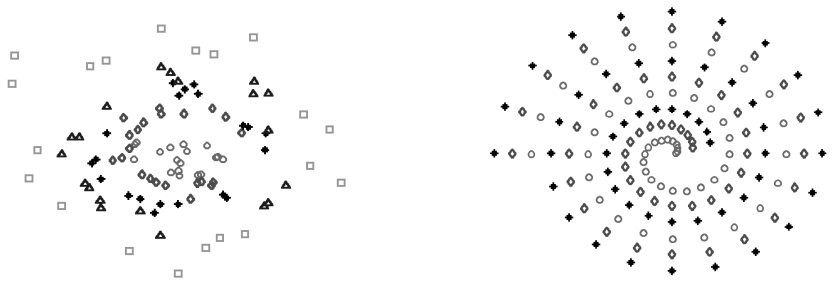
Para el caso linealmente separable, se han construido dos tipos de conjunto de entrenamiento, $\mathcal{T}_1$ y $\mathcal{T}_2$, sobre el plano, $\mathbb{R}^2$, formados por 150 patrones repartidos en 3 clases, con igual número de representantes cada una, generados mediante distribuciones gaussianas de varianza unitaria, Figura 5.1. En el caso de la Figura 5.1(a) los centros de cada distribución han sido situados en línea, mientras que en la Figura 5.1(b) los centros forman triángulo. El uso de ejemplos similares puede ser observado en [Kressel, 1999] para mostrar la eficacia de la clasificación multiclase "por parejas".

Con la intención de reducir el nivel de restricción del problema QP a resolver en la creación de la máquina K-SVCR también se han creado los conjuntos de entrenamiento $\mathcal{T}_3$ y $\mathcal{T}_4$ tomando de forma aleatoria 15 patrones de los 45 de cada una de las tres clases.

El caso no separable será tratado sobre los conjuntos de entrenamiento $\mathcal{T}_5$ y $\mathcal{T}_6$ representados en la Figura 5.2(a) y la Figura 5.2(b), respectivamente. Nuevamente se trata de conjuntos sobre el plano, $\mathbb{R}^2$, con igual número de representantes para cada clase. El conjunto $\mathcal{T}_5$ está formado por 100 patrones repartidos en 5 clases formando una distribución gaussiana y han sido separados en función del radio al centro de esta distribución. Para el conjunto $\mathcal{T}_6$ los 150 puntos de entrenamiento han sido distribuidos formando una espiral.

5.3.2 Factor de insensitividad δ

Las restricciones impuestas sobre los ℓ_3 patrones correspondientes a la etiqueta 0 en la definición de la máquina K-SVCR, tanto para el caso separable (5.7) como para el



(a) Conjunto de entrenamiento $\mathcal{T}_5$. (b) Conjunto de entrenamiento $\mathcal{T}_6$.

Figura 5.2: Conjuntos de entrenamiento no linealmente separables.

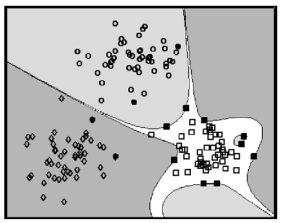
no separable (5.27) provocan que la fisonomía de la función de decisión ternaria (5.8) dependa de forma muy importante de la definición del parámetro de insensitividad δ. En esta Subsección se realizan una serie de experimentos sobre los problemas artificiales definidos anteriormente que permitirán constatar la influencia de este parámetro sobre el resultado final de la clasificación.

Inicialmente se utiliza el conjunto de entrenamiento $\mathcal{T}_1$ para observar el comportamiento de la máquina de aprendizaje en función del factor de insensitividad en un caso separable. Para ello se utilizará una función polinomial de grado 4 como núcleo.

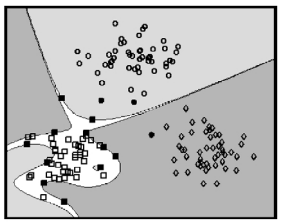
En la Figura 5.3 aparece la evolución de la función de decisión según aumenta el nivel de insensitividad. Las figuras han sido dispuestas en columnas según la elección que se ha determinado de los patrones con etiqueta 0, mientras que a cada fila le corresponde un mismo valor δ. Las cantidades que aparecen debajo de cada subfigura corresponden al factor de insensitividad utilizado, el tiempo de computación y el número de vectores soporte necesarios en la expansión de la función de decisión. Tal como ya fue apuntado, cuanto menor sea el parámetro δ, menor es la generalización del espacio para los patrones con etiqueta 0, mayor es el tiempo de computación y mayor es el número de vectores soporte, los cuales han sido marcados sobre las gráficas.

5.3.3 Comparativa con la SVMR

Utilizando el conjunto de entrenamiento de clases linealmente separables $\mathcal{T}_2$, es entrenada una K-SVCR y una máquina SVMR estándar con salidas $\{-1, 0, +1\}$ asignadas a cada clase en cualquiera de sus combinaciones, por lo que se dispondrá de tres

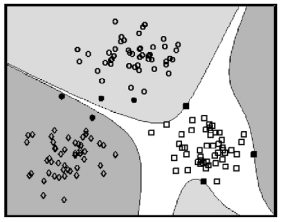


(a) $\varepsilon = 0.050, t = 114.8s, nsv = 13$

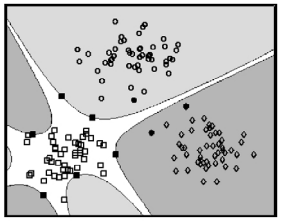


(b) $\varepsilon = 0.050, t = 121.0s, nsv = 14$

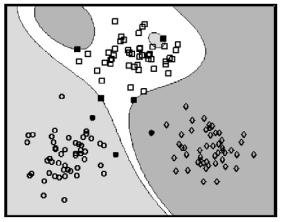
(c) $\varepsilon = 0.050, t = 138.2s, nsv = 14$



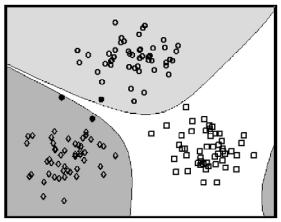
(d) $\varepsilon = 0.250, t = 92.2s, nsv = 7$



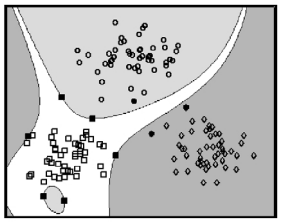
(e) $\varepsilon = 0.250, t = 101.3s, nsv = 9$



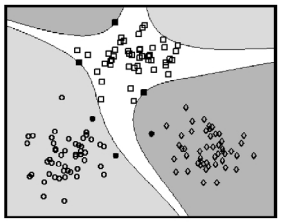
(f) $\varepsilon = 0.250, t = 110.8s, nsv = 7$



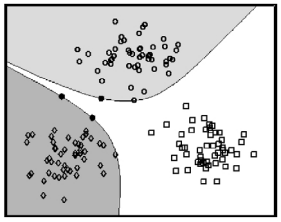
(g) $\varepsilon = 0.500, t = 86.0s, nsv = 3$



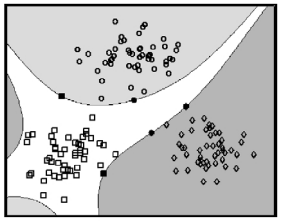
(h) $\varepsilon = 0.500, t = 98.1s, nsv = 9$



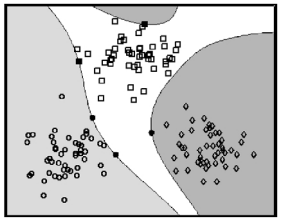
(i) $\varepsilon = 0.500, t = 96.2s, nsv = 6$



(j) $\varepsilon = 0.999, t = 89.1s, nsv = 3$

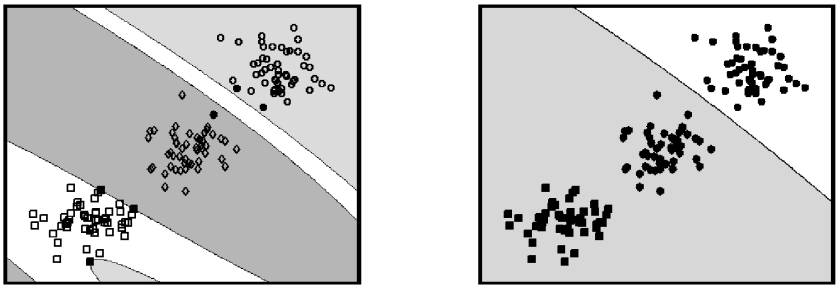


(k) $\varepsilon = 0.999, t = 87.9s, nsv = 5$



(l) $\varepsilon = 0.999, t = 96.8s, nsv = 5$

Figura 5.3: Resultados para diferentes niveles de insensitividad del entrenamiento sobre el conjunto $\mathcal{T}_1$. Las cantidades que acompañan cada subfigura representan el nivel de insensitividad, el tiempo de entrenamiento y el número de vectores soporte utilizados.



(a) Máquina K-SVCR entrenada sobre el conjunto $\mathcal{T}_2$.

(b) Máquina SVMR entrenada sobre el conjunto $\mathcal{T}_2$.

Figura 5.4: Resultados del entrenamiento sobre el conjunto $\mathcal{T}_2$ utilizando funciones núcleo polinomiales de grado $n = 3$.

máquinas para cada tipo de arquitectura. Para el entrenamiento es utilizado un factor $C = \infty$ porque las clases son separables y no se hace necesario el uso de variables artificiales y una insensitividad de nivel medio, $\delta = 0.5$.

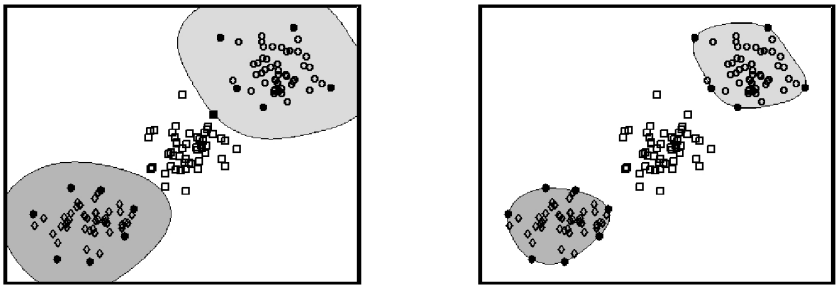
Si se utiliza como núcleo una función polinomial de grado $n = 3$

$$k(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x} \cdot \mathbf{y})^n \ , \tag{5.41}$$

se obtienen los siguientes resultados: las máquinas K-SVCR emplean 93.0, 86.3 y 92.2 segundos en ser entrenadas y la función solución se expande sobre 6, 4 y 5 vectores soporte respectivamente, mientras que la SVMR tras 375.8, 365.9 y 367.6 segundos de entrenamiento, aunque la función solución se expanda sobre los 150 patrones de entrenamiento no consigue clasificar bien como puede apreciarse en las dos gráficas de la Figura 5.4.

El motivo del malfuncionamiento de la SVMR está en la poca amplitud o capacidad del espacio de Hilbert generado por el núcleo polinómico de grado 3, espacio de características $\mathcal{F}$. En el caso K-SVCR el espacio $\mathcal{F}$ resulta suficiente para conseguir separar las clases, pero la mayor exigencia impuesta por las restricciones de la SVMR hacen que el nuevo espacio no sea lo bastante amplio como para permitir la clasificación correcta por esta máquina. Esta experimentación ha permitido ilustrar cómo la exigencia de cubrir todos los patrones de etiqueta 0 por una zona insensitiva entorno a la función de decisión no resulta un requerimiento demasiado extremo, pues incluso una máquina SVMR tradicional implica mayores restricciones.

Para conseguir aumentar la amplitud del espacio de características sobre el conjunto de patrones de entrenamiento existen dos posibilidades, o bien se usa un núcleo que genere un espacio de Hilbert de mayor dimensión o bien se reduce el número de patrones de entrenamiento.



(a) Máquina K-SVCR entrenada sobre el conjunto $\mathcal{T}_2$.

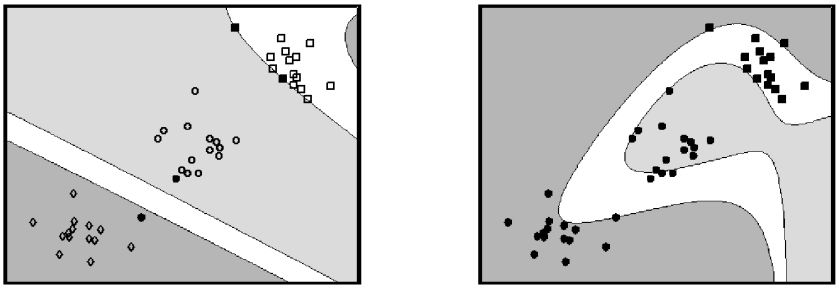
(b) Máquina SVMR entrenada sobre el conjunto $\mathcal{T}_2$.

Figura 5.5: Resultados del entrenamiento sobre el conjunto $\mathcal{T}_2$ utilizando funciones núcleo gaussianas de varianza 0.5.

Para ilustrar la primera posibilidad se ha optado por elegir como funciones núcleo a gaussianas de varianza 0.5. El espacio $\mathcal{F}$ generado por las gaussianas tiene dimensión VC infinita por lo que siempre ha de existir una solución, pero el precio a pagar es un mayor coste en tiempo computacional — no es equivalente evaluar un polinomio que una función exponencial — y la imposibilidad de asegurar generalización. Tras realizar la experimentación se obtienen tres máquinas de aprendizaje K-SVCR que se expanden sobre 12, 13 y 10 vectores soporte y que han necesitado de 102.5, 105.4 y 102.0 segundos respectivamente para ser entrenadas. Nuevamente dos de las SVMRs entrenadas emplean un mayor tiempo computacional en este proceso que la nueva máquina, 495.6 y 480.8 segundos con 12 vectores soporte en ambos casos, mientras que para la tercera de ellas su entrenamiento fue detenido tras más de 14 horas de computación sin haber sido obtenida la solución. En la Figura 5.5 pueden observarse los resultados de dos de las máquinas entrenadas con los dos tipos de arquitectura.

La segunda opción para reducir la demanda en las restricciones de las máquinas de aprendizaje consiste en reducir el número de patrones. Para el siguiente experimento se han seleccionado sólo 15 de los 50 patrones que componen cada clase del conjunto de aprendizaje $\mathcal{T}_1$, conjunto $\mathcal{T}_3$, y se ha entrenado cada tipo de máquina. Como era de esperar, el tiempo de computación es mucho menor y se ha obtenido como resultado 2.2, 1.7 y 1.8 segundos de entrenamiento y 5, 3 y 4 vectores soporte para la nueva máquina, y 6.3, 5.6 y 9.8 segundos para entrenar la máquina SVMR para obtener 7, 2 y 45 vectores soporte. Como puede deducirse de los datos y apreciarse en la Figura 5.6, la tercera de las elecciones de etiquetado vuelve a causar problemas a la máquina SVMR pues incluso expandiendo la solución sobre todos los vectores soporte se cometen errores en el entrenamiento.

Los tres experimentos anteriores — con núcleo polinomial de grado 3 sobre todo el conjunto de entrenamiento, sobre una selección aleatoria de 45 patrones en total, y



(a) Máquina K-SVCR entrenada sobre el conjunto $\mathcal{T}_3$

(b) Máquina SVR entrenada sobre el conjunto $\mathcal{T}_3$

Figura 5.6: Resultados del entrenamiento sobre el conjunto $\mathcal{T}_3$ utilizando funciones núcleo polinomiales de grado $n = 3$.

con núcleo funciones gaussianas de varianza 0.5 sobre todo el conjunto de aprendizaje — han sido también realizados sobre el conjunto $\mathcal{T}_1$ con resultados similares a los obtenidos sobre $\mathcal{T}_2$, como puede observarse en la Tabla 5.1.

5.3.4 Problemas Artificiales No Linealmente Separables

Para ilustrar el funcionamiento de la máquina K-SVCR sobre problemas no linealmente separables se han utilizado los conjuntos de entrenamiento artificiales $\mathcal{T}_5$ y $\mathcal{T}_6$ ya descritos anteriormente.

En el caso del conjunto $\mathcal{T}_5$, se dispone de $K = 5$ clases a ser separadas y de un número muy reducido de patrones para cada clase. Para este problema se ha utilizado como núcleo una función gaussiana que asegure la existencia de solución. El parámetro de insensitividad δ es de 0.75, mayor que el propuesto para los casos linealmente separables, puesto que la zona de 0-etiquetado corresponde ahora a 3 clases y en principio corresponderá con una zona geométrica más amplia que la representada por las dos clases a ser separadas de forma individual, ± 1. Además, el amplio de las campanas de Gauss se ha establecido en $\sigma = 0.45$ intentando que el número de vectores soporte que aparezcan sea más bien elevado en comparación con el número de patrones ya que se dispone de muy pocos de estos últimos.

En la Figura 5.7 se muestran algunas de las máquinas K-SVCR obtenidas en función de las clases a ser separadas y a ser etiquetadas con 0. Los resultados globales han sido sumarizados en la Tabla 5.2, donde se muestran el tiempo de ejecución del entrenamiento en segundos y el número de vectores soporte. Debe resaltarse la

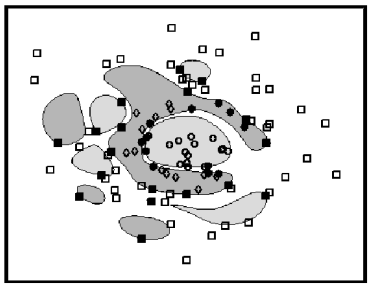
núcleo	parám.	#pat	máquina		1-2-3	1-3-2	2-3-1
polin.	3	150	K-SVCR	tiempo	89.5	96.9	90.5
				#SV	4	8	5
			SVMR	tiempo	366.1	364.4	361.5
				#SV	150	150	150
gauss.	0.5	150	K-SVCR	tiempo	109.1	99.9	102.7
				#SV	15	14	15
			SVMR	tiempo	495.0	470.0	$\geqslant 5400$
				#SV	15	14	?
polin.	3	45	K-SVCR	tiempo	2.3	1.7	2.0
				#SV	4	5	5
			SVMR	tiempo	6.4	5.9	7.2
				#SV	5	5	6

Tabla 5.1: Resultados sobre el conjunto de entrenamiento $\mathcal{T}_1$

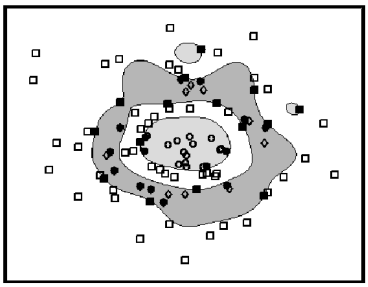
etiquetas	1-2-3,4,5	1-3-2,4,5	1-4-2,3,5	1-5-2,3,4	2-3-1,4,5
tiempo	76.1	47.6	46.2	47.7	63.8
nsv	30	30	31	35	45

etiquetas	2-4-1,3,5	2-5-1,3,4	3-4-1,2,5	3-5-1,3,4	4-5-1,2,3
tiempo	54.9	58.0	118.1	58.3	60.5
nsv	36	44	48	38	38

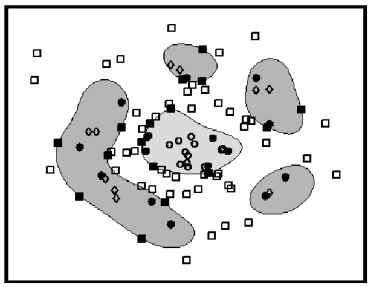
Tabla 5.2: Resultados sobre el conjunto de entrenamiento $\mathcal{T}_5$ utilizando núcleos gaussianos.



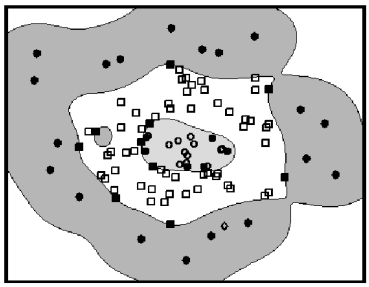
(a) $1-2-3,4,5$



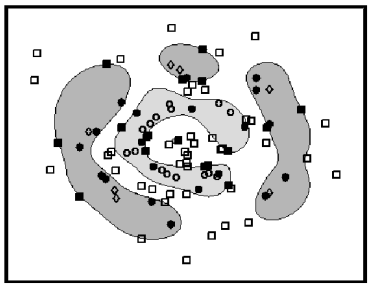
(b) $1-3-2,4,5$



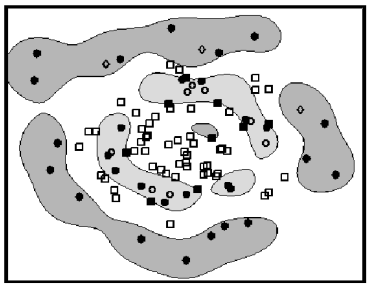
(c) $1-4-2,3,5$



(d) $1-5-2,3,4$



(e) $2-4-1,3,5$



(f) $3-5-1,2,4$

Figura 5.7: Resultados del entrenamiento sobre el conjunto $\mathcal{T}_5$ utilizando núcleos gaussianos para diferentes elecciones de etiqueta de clase.

etiquetas	1-2-3,4,5	1-3-2,4,5	1-4-2,3,5	1-5-2,3,4	2-3-1,4,5
tiempo	181.8	79.1	101.5	71.9	114.7
nsv	64	50	48	33	64

etiquetas	2-4-1,3,5	2-5-1,3,4	3-4-1,2,5	3-5-1,3,4	4-5-1,2,3
tiempo	70.5	31.6	78.3	65.4	54.1
nsv	44	100	48	31	20

Tabla 5.3: Resultados sobre el conjunto de entrenamiento $\mathcal{T}_5$ utilizando núcleos polinomiales.

etiquetas	1-2-3	1-3-2	2-3-1
tiempo	244.8	222.6	316.3
nsv	98	79	95

Tabla 5.4: Resultados sobre el conjunto de entrenamiento $\mathcal{T}_6$ utilizando núcleos gaussianos.

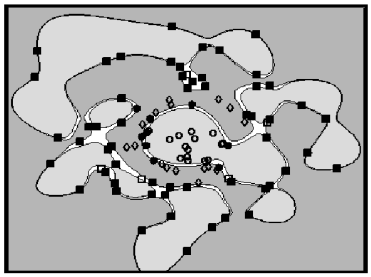
especial dificultad que le comporta a la máquina realizar la separación cuando las clases no etiquetadas nulas se hallan contiguas.

También ha sido utilizado un entrenamiento sobre núcleos polinomiales de grado 10 con igual factor de insensitividad que en el caso anterior, $\delta = 0.75$, con la intención de mostrar la eficiencia de la nueva máquina sobre un espacio de dimensión VC finita. Los resultados han sido reflejados en las gráficas de la Figura 5.8 y la Tabla 5.3.

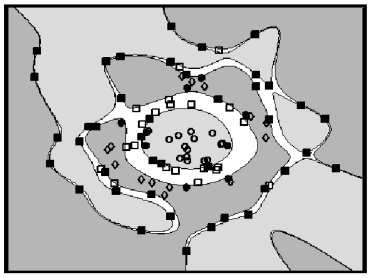
Para el conjunto de entrenamiento $\mathcal{T}_6$ se vuelve a tener 3 clases con los patrones repartidos equitativamente formando espirales anidadas. Utilizando núcleos gaussianos con los parámetros habituales $\sigma = 0.5$ y $\delta = 0.5$ se obtienen los resultados mostrados en la Figura 5.9 y recogidos en la Tabla 5.4.

5.4 En Resumen

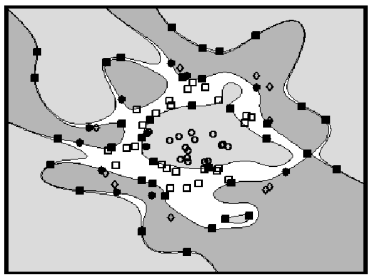
En el presente Capítulo se ha presentado e ilustrado el funcionamiento de un nuevo algoritmo basado en vectores soporte con el objetivo posterior de ser utilizado durante el esquema de descomposición de multiclasificación, denominado algoritmo K-SVCR. El problema implícito motivador de la presente formulación ha sido la mala disposición de las clases cuando se utiliza las metodologías estándares de una contra una o una contra el resto de clases. Al tratarse de un problema de clasificación multiclase, $K > 2$, toma sentido construir máquinas de aprendizaje que asignen salida $+1$ o -1 si el patrón de entrenamiento pertenece a las clases a ser separadas, y salida 0 si por contra el patrón tiene una etiqueta diferente a las anteriores, forzando al hiperplano



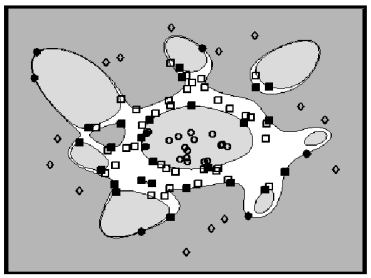
(a) $1-2-3,4,5$



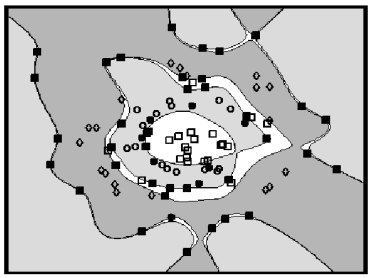
(b) $1-3-2,4,5$



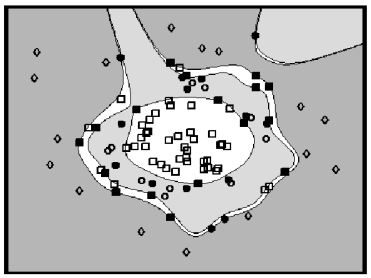
(c) $1-4-2,3,5$



(d) $1-5-2,3,4$

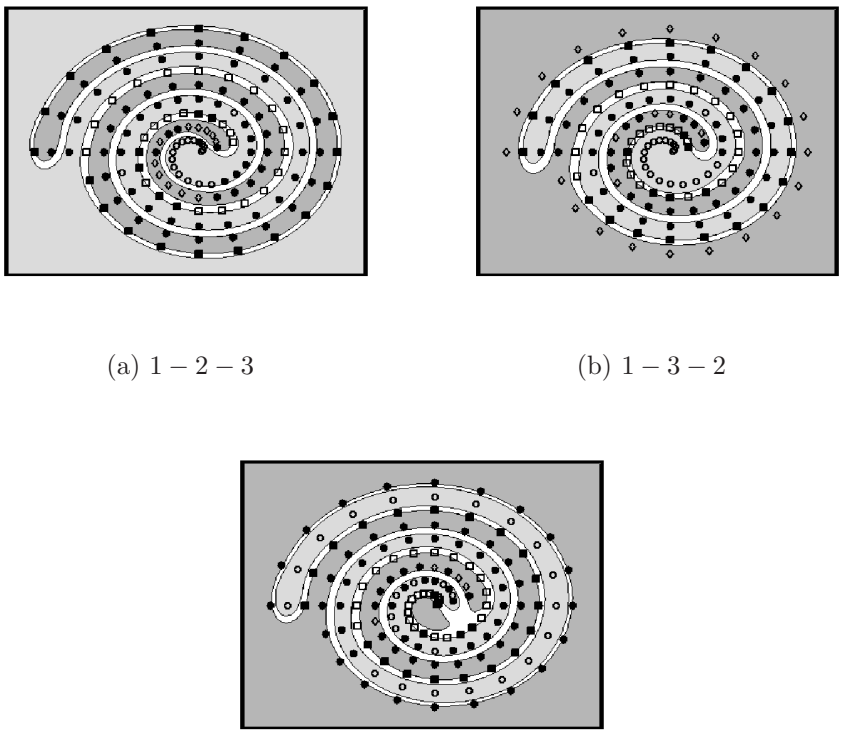


(e) $2-4-1,3,5$



(f) $3-5-1,2,4$

Figura 5.8: Resultados del entrenamiento sobre el conjunto $\mathcal{T}_5$ utilizando núcleos polinomiales para diferentes elecciones de etiqueta de clase.



(a) $1-2-3$ (b) $1-3-2$

(c) $2-3-1$

Figura 5.9: Resultados del entrenamiento sobre el conjunto $\mathcal{T}_6$ utilizando núcleos gaussianos para diferentes elecciones de etiqueta de clase.

separador de las dos clases implicadas a recubrir todos los patrones de entrenamiento pertenecientes a cualquier otra clase que no sean las dos iniciales.

Se ha desarrollado la formulación matemática del problema QP asociado tanto para el caso linealmente separable como para el no linealmente separable. Este problema de programación cuadrática podría ser interpretado como un intermedio entre el método SVMC y el método SVMR de entrenamiento sobre vectores soporte para estimación de regresiones.

Finalmente, se ha mostrado el modo de funcionamiento de las K-SVCRs aplicado sobre problemas artificiales que permitan su fácil visualización. Así, se ha mostrado su eficiencia respecto a una SVMR con tres salidas gracias a su menor demanda en las restricciones del problema QP asociado; el nuevo factor de insensitividad ligado a aquel definido por Vapnik para crear las SVMRs resulta crítico en la definición de la tercera clase, su amplitud y el número de vectores soporte; la formulación permite trabajar con eficiencia con cualquier número de clases tal cómo se ha podido observar con un problema no linealmente separable de $K = 5$ clases.