Métodos de combinación de clasificadores sobre Máquinas de Vectores Soporte*

Juan José Rodríguez
Lenguajes y Sistemas Informáticos
Escuela Politécnica Superior
Universidad de Burgos
jjrodriguez@ubu.es

Oscar J. Prieto, Carlos J. Alonso
Departamento de Informática
ETS Ingeniería Informática
Universidad de Valladolid
{oscapri,calonso}@infor.uva.es

Resumen

En este trabajo se presentan los resultados experimentales obtenidos al aplicar métodos de combinación de clasificadores homogéneos (en los que todos los clasificadores de la combinación se han obtenido utilizando un mismo método) sobre Máquinas de Vectores Soporte (SVM). Los métodos de combinación considerados son Bagging, Boosting y el basado en Rotaciones. Se considerán los kernel lineal y cuadrático. En el segundo caso, el método de combinación que mejor funciona es el basado en Rotaciones.

1. Introducción

En este trabajo se estudia la aplicación de los métodos de combinación de clasificadores Bagging [2], Boosting [9] y basado en Rotaciones [8] sobre SVM con kernel lineal y cuadrático.

Otro trabajo en el que se consideran métodos de combinación de clasificadores sobre SVM es [10], en el que se proponen nuevos métodos, basados en el análisis de la descomposición bias-varianza del error. Considera una variante de Bagging, en la que los clasificadores base se ajustan con el objetivo de tener un bias bajo, ya que Bagging es un método que reduce la varianza. En el presente trabajo no se ajustan parámetros

*Este trabajo ha sido financiado por el proyecto del MCyT DPI2002–01809.

El resto del trabajo se organiza como sigue. El método de combinación basado en rotaciones se describe en la sección 2. En la sección 3 se discute sobre la adecuación de este método de combinación de clasificadores para su uso con SVM. La sección 4 está dedicada a la validación experimental. Finalmente, la sección 5 presenta las conclusiones y trabajos futuros.

2. Combinaciones basadas en Rotaciones

El método ya se presentó en [8], se vuelve a presentar aqui para conveniencia del lector. Se basa en transformar el conjunto de datos y aplicar el método de clasificación base sobre el conjunto de datos transformado. Los resultados de los distintos clasificadores se combinan por votación.

El proceso de transformación aquí considerado se basa en el Análisis de Componentes Principales (PCA) [4]. Las componentes son combinaciones lineales de los atributos. Normalmente esta técnica se utiliza para reducir la dimensionalidad del conjunto de datos, pues las componentes aparecen ordenadas por su importancia, así que se pueden descartar las menos útiles. No obstante, aquí se van a utilizar todas las componentes, puesto que en ese caso se obtiene un conjunto de datos diferente con exactamente la misma información que el original, pero en el que se ha producido una rotación de ejes. Esta técnica trabaja con atributos numéricos, así que si el conjunto de da-

Actas del III Taller Nacional de Minería de Datos y Aprendizaje, TAMIDA2005, pp.185-192
ISBN: 84-9732-449-8

tos contiene atributos discretos, éstos deberán transformarse en numéricos.

Uno de los objetivos que se busca cuando se combinan clasificadores es que estos sean diversos, pues si son iguales no se gana con su combinación. El otro, que es contradictorio con el anterior, es que los clasificadores base sean precisos. Entendemos que el análisis de componentes principales es adecuado como mecanismo de transformación del conjunto de datos porque por un lado no se pierde información, y la precisión del clasificador no se ve afectada por dicha perdida como puede ocurrir cuando se utiliza Bagging, pero por otro lado el modelo generado por el clasificador puede ser bastante diferente, siempre y cuando el método de inducción no sea invariante a las rotaciones.

Ahora bien, dado un conjunto de datos, el resultado de aplicar sobre el mismo el análisis de componentes principales es único. Para su uso en un sistema de combinación de clasificadores necesitamos diferentes versiones del conjunto de datos. Para ello, las posibles opciones son:

- Aplicar el análisis a un subconjunto de las instancias. Salvo que el número de instancias seleccionado sea pequeño, los resultados serán bastante parecidos.
- Aplicar el análisis a un subconjunto de las clases. En realidad, este caso es una variante del anterior, pues al seleccionar clases se están seleccionando instancias. Sin embargo, es de suponer que se obtendrán resultados más diferentes si se consideran las clases por separado.
- Aplicar el análisis a un subconjunto de los atributos. En este caso solamente se rotan los atributos en cuestión. Con eso solamente los atributos del subconjunto se ven modificados. Para modificar todos los atributos, lo que se puede hacer es agrupar los atributos en grupos, y aplicar el análisis de manera independiente en cada uno de ellos.

Hay que tener presente que las estrategias anteriores se tienen en cuenta sólo para aplicar el análisis, pero que una vez que se han obtenido las componentes sobre el subconjunto de datos, entonces todo el conjunto de datos se reformula utilizando dichas componentes. Combinando las estrategias anteriores, es posible, salvo para conjuntos de datos muy pequeños, obtener tantos conjuntos de datos diferentes como sean necesarios.

El aplicar el análisis en subgrupos de atributos, además de ser otro mecanismo de variación, proporciona ventajas adicionales. Por un lado, el tiempo de ejecución del análisis depende principalmente del número de atributos, por lo que es bastante más rápido hacer el análisis en subgrupos que hacer el análisis con todos los atributos.

Por otro lado, se reducen notablemente los requisitos de almacenamiento. Una componente no es más que una combinación lineal de los atributos originales, si se mantienen todas se necesita una matriz para almacenar todos los coeficientes, por lo que el espacio requerido es cuadrático en el número de atributos cada vez que se realiza el análisis. Si el análisis se hace en grupos, el número de coeficientes a almacenar es cuadrático en el número de atributos por grupo, y lineal en el número de atributos totales.

Un parámetro adicional del método es un coeficiente c que indica la relación entre el número de atributos del conjunto de datos original y del final. Por ejemplo, si $c = 0{,}5$, el conjunto de datos resultado tendría la mitad de atributos que el original. En el presente trabajo $c = 1$.

Una descripción algorítmica del método se presenta en la figura 1.

3. Combinaciones basadas en Rotaciones y SVM

El método de combinación de clasificadores basado en rotaciones se introdujo para aprovechar una de las características de los árboles de decisión, el que las superficies obtenidas están delimitadas por paralelas a los ejes.

Al utilizar SVM esa limitación no existe, por lo que la principal causa de diversidad desaparece. Sin embargo, se pueden obtener clasifica-

Entrada:
- C: conjunto de datos
- s: entero (tamaño de los grupos)
- c: real (coeficiente entre el nº de atributos final y el inicial)

Algoritmo:
- $R \leftarrow$ conjunto de datos vacío
- $n \leftarrow$ nº de atributos en C (excluyendo la clase)
- utilizado$[1 \ldots n] \leftarrow$ Falso
- $g \leftarrow \lceil c \times n/s \rceil$ (nº de grupos)
- Para $i \leftarrow 1 \ldots g$ hacer
  - $A \leftarrow$ conjunto de datos vacío
  - Para $j \leftarrow 1 \ldots s$ hacer
    - Si existe un k tal que utilizado$[k]$ es Falso
      - Seleccionar aleatoriamente un valor de k tal que utilizado$[k]$ sea Falso y el atributo a_k de C no pertenezca a A
    - En otro caso
      - Seleccionar aleatoriamente un valor de k tal que el atributo a_k de C no pertenezca a A
    - $A \leftarrow A \cup \{a_k\}$
  - $P \leftarrow$ análisis de la componentes principales de un subconjunto de las instancias de A
  - $A' \leftarrow$ transformación de A al espacio de coordenadas definido por P
  - $R \leftarrow R \cup A'$
- $R \leftarrow R \cup \{C.\text{clase}\}$

Salida: R

Figura 1: Descripción del método para la transformación del conjunto de datos.

dores diversos al utilizar SVM por:

- La normalización. La implementación de SVM utilizada normaliza los atributos. Es decir, los atributos rotados, se normalizan. Esto altera las superficies de decisión.

- El número de atributos. El método basado en rotaciones permite especificar cuantos atributos se desean en el conjunto de datos generado. En el presente trabajo este número es similar al número de atributos originales, pero no idéntico, ya que tiene que ser múltiplo del número de atributos por grupo. Por ejemplo, si hay 4 atributos que se agrupan de 3 en 3, el conjunto resultante tendrá 6 atributos. Si un atributo se selecciona más de una vez, esto altera las superficies de decisión.

- El uso de kernels no lineales.

La figura 2 muestra las superficies de decisión obtenidas para el conjunto de datos bidimensional denominado "conus-torus", introducido en [5]. Se muestran las superficies obtenidas con SVM sobre los datos originales y sobre modificaciones de los datos. Por un lado, el remuestreo, que es la transformación que se hace con los datos en Bagging. Por otro, el aplicar rotaciones.

Al utilizar el kernel lineal todas las superficies son bastante similares, aunque no idénticas. Al utilizar el kernel cuadrático, se obtienen superficies más diferentes.

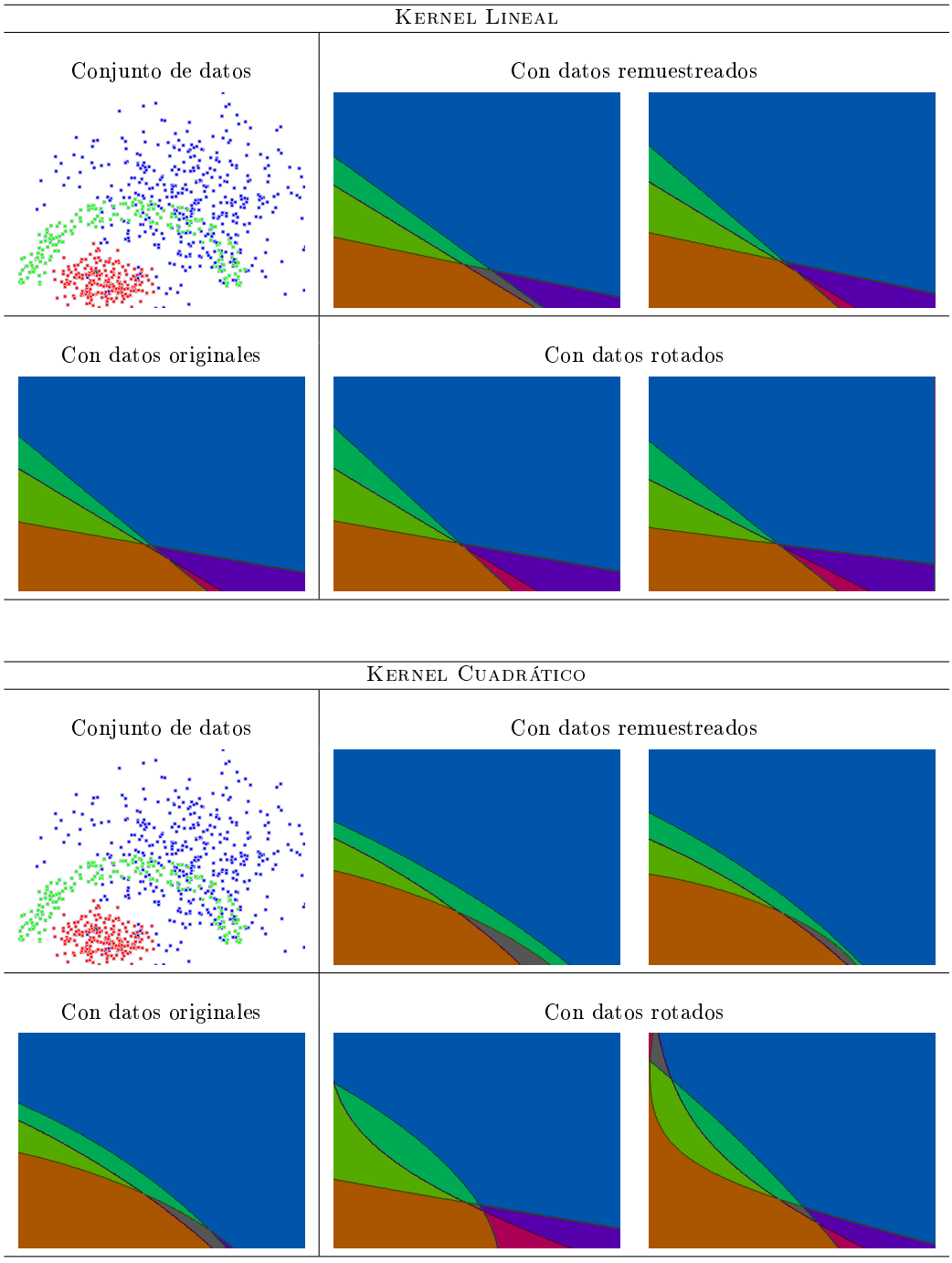

Figura 2: Superficies de decisión obtenidas para el conjunto de datos "conus-torus".

4. Validación experimental

4.1. Configuración

Los métodos utilizados fueron SVM, Bagging de SVM, Boosting (AdaBoost.M1) de SVM y Rotaciones sobre SVM. Las implementaciones utilizadas, tanto de los clasificadores base como de los métodos de combinación fueron las disponibles en WEKA [11], salvo el método de Rotaciones, implementado por los autores. El número de clasificadores a combinar fue 10, ya que los clasificadores base son bastante complejos.

En el caso de las combinaciones basadas en rotaciones los atributos se agruparon de 3 en 3. El número de atributos generados era aproximadamente el mismo que había en el conjunto de datos original. Cada atributo nominal se convirtió en tantos atributos binarios como posibles valores.

Para cada método y conjunto de datos se realizaron 15 validaciones cruzadas, cada una de ellas con 10 grupos. Para comparar los métodos se utilizó la implementación en WEKA del *estadístico test-t remuestreado corregido* propuesto en [7].

4.2. Conjuntos de datos

Los conjuntos de datos utilizados son los que aparecen en el cuadro 1. Todos ellos tienen su origen en el repositorio de la UCI [1]. Algunos conjuntos de datos se modificaron ligeramente. Por un lado, tanto en "splice" como en "zoo" se eliminó un atributo que identificaba las instancias. Este tipo de atributos no son útiles para el aprendizaje, y en la implementación actual, en la que se convierten los atributos discretos en numéricos origina una sobrecarga notable, pues se obtienen tantos nuevos atributos como instancias. Por otro lado, el conjunto de datos "vowel" incluía un atributo indicando si el ejemplo era para entrenamiento o para prueba, que obviamente no puede aparecer dentro del conjunto de datos. Además, en este conjunto de datos se incluye información sobre quién es el locutor y cuál es su sexo. Hay autores que incluyen estos dos atributos como parte del conjunto de datos, pero también hay trabajos en los que no se consideran esos atributos. Por eso se consideran dos versiones, "vowel-c" que incluye esta información de contexto y "vowel-n" que no la incluye.

Cuadro 1: Características de los conjuntos de datos.

conjunto de datos	clases	ejemplos	discretos	contínuos
anneal	6	898	32	6
audiology	24	226	69	0
autos	7	205	10	16
balance-scale	3	625	0	4
breast-cancer	2	286	10	0
cleveland-14-heart	5	307	7	6
credit-rating	2	690	9	6
german-credit	2	1000	13	7
glass	7	214	0	9
heart-statlog	2	270	0	13
hepatitis	2	155	13	6
horse-colic	2	368	16	7
hungarian-14-heart	5	294	7	6
hypothyroid	4	3772	22	7
ionosphere	2	351	0	34
iris	3	150	0	4
kr-vs-kp	2	3196	37	0
labor	2	57	8	8
lymphography	4	148	15	3
mushroom	2	8124	22	0
pima-diabetes	2	768	0	8
primary-tumor	22	239	17	0
segment	7	2310	0	19
sonar	2	208	0	60
soybean	19	683	35	0
splice	3	3190	60	0
vehicle	4	846	0	18
vote	2	435	16	0
vowel-c	11	990	2	10
vowel-n	11	990	0	10
waveform	3	5000	0	40
wisconsin-breast	2	699	0	9
zoo	7	101	16	2

4.3. Resultados

El cuadro 2 muestra los resultados obtenidos para kernel lineal. Se compara el método de las

rotaciones con el resto de métodos. En el cuadro 3.a se resumen los resultados, indicando cuántas veces un método es mejor que otro.

Estos resultados son bastante desfavorables para el método de las rotaciones. Comparado con utilizar un solo clasificador hay una victoria y una derrota. Comparado con boosting, hay 2 victorias y 5 derrotas. Es razonable este pobre comportamiento del método, ya que al ser los clasificadores lineales la rotación de los datos dará lugar a clasificadores muy parecidos. Para que una combinación tenga éxito es necesario que los clasificadores a combinar sean diversos entre si [6].

Cuadro 3: Resumen de resultados. Se muestra el número de conjuntos de datos para los que el método de la columna va mejor que el de la fila. Entre paréntesis se indica en cuántos casos las diferencias son significativas.

a) Kernel lineal

	SVM	Bag.	Boost.	Rot.
SVM	-	19 (1)	13 (4)	16 (1)
Bagging	13 (0)	-	15 (4)	15 (1)
Boosting	15 (1)	17 (2)	-	19 (2)
Rotaciones	16 (1)	16 (2)	13 (5)	-

b) Kernel cuadrático

	SVM	Bag.	Boost.	Rot.
SVM	-	19 (2)	14 (4)	25 (4)
Bagging	13 (1)	-	13 (5)	22 (4)
Boosting	12 (3)	19 (4)	-	22 (7)
Rotaciones	7 (2)	10 (1)	10 (4)	-

Los resultados con kernel cuadrático se muestran en el cuadro 4, mientras que la relación entre métodos se resume en el cuadro 3.b. En este caso claramente el mejor método es el de las rotaciones. Sin embargo las diferencias entre métodos son escasas, al comparar con el uso de un solo clasificador el resultado es de 4 a 2, sobre 33 conjuntos de datos.

5. Conclusiones y trabajos futuros

Si bien las ventajas de utilizar métodos estándar de combinación de clasificadores no son tan aparentes al utilizar SVM como al utilizar otros métodos de clasificación más inestables, como los árboles de decisión, es todavía posible obtener mejoras en la precisión.

De los métodos de combinación considerados, al utilizar kernel cuadrático, el que mejores resultados proporciona es el basado en rotaciones. Por ejemplo, de 33 conjuntos de datos en 25 de ellos se obtienen mejores resultados con una combinación basada en rotaciones que con un solo clasificador. Ahora bien, al exigir que esas diferencias sean significativas el resultado se reduce a 4 frente a 2.

Al utilizar kernel lineal el método de combinación basado en rotaciones no proporciona buenos resultados, lo que era de esperar ya que los clasificadores son lineales. En consecuencia, los miembros de la combinación serán demasiado semejantes entre si.

En el presente trabajo se han considerado sólo los kernels lineal y cuadrático. El otro utilizado con más frecuencia es el kernel gaussiano. Por tanto, se plantea como trabajo futuro el utilizar este tipo de kernels. El motivo de que esta tarea no se haya abordado todavía es que el funcionamiento de SVM con este tipo de kernels es bastante sensible a los valores de los parámetros (i.e., C y δ). La herramienta utilizada, WEKA, no proporciona un mecanismo para seleccionar estos valores. LIBSVM realiza automáticamente esta selección utilizando validación cruzada, pero no dispone de métodos de combinación de clasificadores.

Por otro lado, el seleccionar los parametros para cada miembro de la combinación puede requerir un tiempo excesivo. Una posibilidad a explorar sería seleccionar los parámetros sólo para el primer miembro de la combinación, y utilizar esos valores para el resto de miembros, con la esperanza de que buenos valores de los parámetros para un clasificador también lo sean para el resto.

Otro aspecto a considerar es el tratamiento de conjuntos de datos con más de dos clases. Dado que los métodos de combinación de clasificadores considerados pueden trabajar combinando clasificadores multiclase, y el método utilizado para obtener los clasificadores admite conjuntos multiclase (utilizando "1 contra 1"), se ha seguido la opción más natural, combinar clasificadores multiclase. Pero otra opción sería utilizar el método "1 contra 1" (u otro que permita obtener un clasificador multiclase a

Cuadro 2: Resultados con kernel lineal

Conjunto	Rotaciones-SVM	SVM	Bagging-SVM	Boosting-SVM
anneal	96.47±1.78	97.42±1.53	97.71±1.50 ○	99.45±0.73 ○
audiology	82.49±6.72	80.45±6.68	78.21±6.14 ●	81.46±6.99
autos	70.16±9.74	70.81±9.86	70.67±9.48	74.34±9.47
balance-scale	86.74±2.75	87.50±2.56	87.35±2.73	87.49±2.57
breast-cancer	70.27±7.58	69.57±7.54	69.22±6.47	68.28±7.80
cleveland-14-heart	84.10±6.44	83.90±6.19	83.88±6.38	83.70±6.22
credit-rating	85.21±4.08	84.85±4.00	85.04±4.08	83.61±4.05
german-credit	75.27±3.53	75.17±3.58	75.27±3.35	75.16±3.64
glass	57.44±8.70	57.50±8.65	57.57±9.52	59.08±8.96
heart-statlog	83.70±6.28	83.75±6.13	83.93±5.93	83.63±6.19
hepatitis	85.67±8.41	85.86±8.64	85.83±8.31	82.54±7.93
horse-colic	83.82±5.49	82.81±5.61	82.95±5.80	78.48±7.22 ●
hungarian-14-heart	82.89±6.69	83.09±6.41	83.62±6.18	83.62±5.69
hypothyroid	93.55±0.43	93.58±0.44	93.59±0.42	94.76±0.67 ○
ionosphere	87.43±5.03	88.05±5.33	88.47±5.03	89.29±4.99
iris	96.44±4.74	96.22±4.72	96.13±4.90	97.56±4.26
kr-vs-kp	93.30±3.87	95.81±1.28 ○	96.03±1.20 ○	97.13±0.95 ○
labor	91.51±11.04	92.42±10.65	92.18±11.36	89.73±12.40
lymphography	85.83±8.54	86.37±7.92	85.52±7.99	83.44±9.58
mushroom	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00
pima-diabetes	76.95±4.40	76.83±4.45	77.13±4.43	76.83±4.45
primary-tumor	48.10±6.03	47.38±6.38	47.98±6.36	47.38±6.38
segment	93.27±1.46	92.92±1.46	92.86±1.50	92.89±1.49
sonar	77.85±8.23	77.11±8.37	77.66±8.84	78.22±9.01
soybean	93.56±2.53	93.05±2.64	93.22±2.74	92.85±3.00
splice	94.35±1.18	92.94±1.36 ●	94.23±1.21	92.90±1.31 ●
vehicle	75.10±3.97	74.11±3.99	74.32±3.80	74.13±4.04
vote	95.48±2.81	95.91±2.86	95.89±2.86	95.38±3.20
vowel-c	70.34±4.22	70.96±4.19	71.63±4.19	78.43±3.94 ○
vowel-n	69.16±4.43	68.65±4.54	67.70±4.74	68.65±4.56
waveform	86.50±1.48	86.44±1.47	86.58±1.40	86.44±1.47
wisconsin-breast-cancer	96.71±2.04	96.74±2.02	96.70±2.03	96.72±2.04
zoo	92.34±7.47	93.46±6.97	93.61±7.44	96.61±5.64 ○

○, ● mejora o empeoramiento significativos

partir de clasificadores binarios, como "1 contra todos" o los ECOC [3]) en conjunción con los métodos de combinación de clasificadores sobre problemas binarios.

Referencias

[1] C.L. Blake and C.J. Merz. UCI repository of machine learning databases, 1998. `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

[2] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[3] Thomas G. Dietterich and Ghulum Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1995.

[4] Jiawei Han and Micheline Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2001.

[5] Ludmila Kuncheva and Christopher J. Whitaker. Using diversity with three variants of boosting: Aggressive, conservative, and inverse. In *Multiple Classifier Systems 2002*, pages 81–90, 2002.

[6] Ludmila I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.

[7] Claude Nadeau and Yoshua Bengio. Inference for the generalization error. *Machine Learning*, 52(239–281), 2003.

[8] Juan José Rodríguez and Carlos J. Alonso. Rotation-based ensembles. In *Current Topics in Artificial Intelligence: 10th Conference of the Spanish Association for Artificial Intelligence*, volume

Cuadro 4: Resultados con kernel cuadrático

Conjunto	Rotaciones-SVM	SVM	Bagging-SVM	Boosting-SVM
anneal	99.64±0.62	99.55±0.67	99.33±0.82	99.28±0.80
audiology	82.42±6.76	81.04±6.45	77.50±5.81 •	81.04±6.45
autos	79.37±8.50	77.03±9.07	77.36±8.99	77.03±9.07
balance-scale	90.84±1.48	91.38±1.91	91.32±1.92	93.94±2.98 ○
breast-cancer	72.38±6.97	66.57±8.34 •	67.51±8.01	65.47±7.66 •
cleveland-14-heart	83.93±5.80	81.26±5.92	80.82±5.96 •	77.96±6.71 •
credit-rating	85.96±3.97	84.99±4.11	85.54±3.99	81.88±4.33 •
german-credit	74.41±4.04	69.27±4.46 •	72.52±4.03	69.00±4.34 •
glass	64.02±8.60	62.34±8.84	63.63±9.14	62.62±9.10
heart-statlog	82.91±6.13	81.46±6.43	81.78±6.56	78.74±8.11
hepatitis	83.18±8.22	82.44±8.07	84.23±8.81	83.96±8.21
horse-colic	79.57±6.44	76.74±6.60	80.31±6.76	78.05±7.47
hungarian-14-heart	83.25±5.75	82.82±6.13	82.78±5.64	79.76±6.49
hypothyroid	94.69±0.60	93.62±0.56 •	93.64±0.56 •	94.82±0.81
ionosphere	91.18±4.32	90.77±4.44	91.00±4.33	87.79±5.04 •
iris	96.98±4.13	95.60±4.81	95.64±5.18	96.27±4.60
kr-vs-kp	98.59±2.16	99.47±0.38	99.41±0.46	99.48±0.38
lymphography	84.36±9.11	83.36±8.62	84.63±7.96	83.36±8.62
labor	92.11±11.56	92.80±11.49	91.38±12.46	92.80±11.49
mushroom	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00
pima-diabetes	77.06±4.18	77.33±4.29	77.26±4.07	77.29±4.31
primary-tumor	43.27±6.79	41.52±6.41	44.50±6.48	40.95±6.27
segment	95.93±1.30	94.66±1.46 •	94.69±1.44 •	95.42±1.40
sonar	84.55±8.02	84.14±7.37	84.05±7.73	84.14±7.37
soybean	93.63±2.69	92.81±2.60	93.07±2.61	92.81±2.59
splice	94.41±1.23	96.41±0.93 ○	96.46±0.94 ○	96.42±0.94 ○
vehicle	81.52±3.69	80.39±3.92	80.42±3.70	80.86±3.98
vote	95.75±2.84	94.37±3.38	95.46±3.09	94.96±3.22
vowel-c	96.26±1.86	97.30±1.56 ○	97.23±1.60	98.38±1.35 ○
vowel-n	85.15±3.24	84.25±3.34	84.46±3.46	89.99±3.01 ○
waveform	86.10±1.41	86.02±1.47	86.01±1.41	83.06±1.59 •
wisconsin-breast-cancer	96.72±2.00	96.41±2.20	96.37±2.32	95.48±2.50 •
zoo	95.13±6.64	96.11±5.72	95.65±6.23	96.11±5.72

○, • mejora o empeoramiento significativos

3040 of *Lecture Notes in Artificial Intelligence*, pages 498–506. Springer, 2004.

[9] Robert E. Schapire. The boosting approach to machine learning: An overview. In *MSRI Workshop on Nonlinear Estimation and Classification*, 2002. `http://www.cs.princeton.edu/~schapire/papers/msri.ps.gz`.

[10] Giorgio Valentini. *Ensemble methods based on bias-variance analysis*. PhD thesis, UniversitÃ di Genevoa, 2003.

[11] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999. `http://www.cs.waikato.ac.nz/ml/weka`.