Un estudio empírico preliminar sobre los tests estadísticos más habituales en el aprendizaje automático.

Francisco Herrera[1], Cesar Hervás[2], José Otero[3], and Luciano Sánchez[3]

[1] Depto de CCIA. ETSI Informática. Universidad de Granada. 18071 Granada
[2] Depto. de Informática y A. N. Campus de Rabanales. U. Córdoba. 14071 Córdoba
[3] Depto. de Informática. Campus de Viesques. Universidad de Oviedo. 33204 Gijón

Resumen Actualmente no existe un diseño experimental que sea admitido de forma universal por los investigadores en aprendizaje automático. Hay opiniones diversas en lo referente a la proporción de ejemplos de la muestra que se debe reservar para la fase de validación, o acerca de la forma en que se deben seleccionar estos ejemplos, por mencionar algunos puntos controvertidos. En este trabajo se revisa la bibliografía más relevante al respecto, y se discuten las conclusiones preliminares obtenidas mediante un análisis empírico de la potencia de varios tests, usados comúnmente por los investigadores en minería de datos. El estudio experimental se instrumenta sobre varios conjuntos de datos sintéticos, con propiedades teóricas conocidas.

1. Introducción

Existen distintos factores que hacen necesario emplear algún tipo de test estadístico cuando se evalúan o se comparan algoritmos de aprendizaje. Según [12], estos factores incluyen la métrica del error, la elección de los conjuntos de entrenamiento y test, y la propia naturaleza del algoritmo, cuando este no es determinista.

En este trabajo, un experimento consiste en resolver una serie de problemas usando una implementación de un algoritmo. El conjunto de problemas, medidas realizadas, los detalles de la implementación y, en general, el contexto que acompaña a la realización de los experimentos, y que puede ser relevante de cara a la extracción de conclusiones sobre las medidas realizadas, conforma el *diseño experimental* utilizado [7][18].

La elección de un diseño experimental adecuado para un problema de aprendizaje automático es un punto de controversia entre la comunidad científica [12][25][16][32]. En trabajos recientes, como [31], los algoritmos de aprendizaje se evalúan mediante la comparación de sus resultados sobre conjuntos de datos conocidos [3], utilizando un test estadístico para juzgar la relevancia de las diferencias. Este mismo enfoque será seguido en este trabajo, si bien somos conscientes de que algunos autores cuestionan el que sea posible extraer conclusiones sobre el rendimiento de un algoritmo utilizando los conjuntos de ejemplos

más habituales [15][28][34], y que, por otra parte, la naturaleza de estos diseños experimentales es tal que frecuentemente se vulneran una o más de las condiciones que han de cumplirse para la aplicación de determinado test estadístico [24][8][25].

Este capítulo está organizado en dos partes. En la primera, se realiza una taxonomía de los diseños experimentales más frecuentes que se utilizan en aprendizaje automático. En la segunda, se realizará un estudio empírico de una selección de estos diseños experimentales sobre un problema sintético, de solución conocida, y se extraerán conclusiones sobre la potencia de los tests estadísticos más frecuentes.

2. Diseños experimentales más habituales

2.1. Validación cruzada

La validación cruzada [29][30] es el diseño experimental más utilizado entre los investigadores en aprendizaje automático. En este método, los datos disponibles se dividen aleatoriamente en un conjunto de entrenamiento y un conjunto de test. El conjunto de entrenamiento se subdivide, a su vez, en dos conjuntos disjuntos

- El *conjunto de estimación*, usando para seleccionar el algoritmo.
- El *conjunto de validación*, usando para probar o validar el algoritmo.

La motivación de esta división está en validar el algoritmo sobre un conjunto de datos diferente del empleado para estimar sus parámetros.

Existen numerosas variantes de la validación cruzada. La que se se ha mencionado es conocida como el método *hold out*, y es menos utilizada en la actualidad que la *multifold cross validation* o *k-fold cross validation*. Esta última consiste en dividir el conjunto de ejemplos de que se dispone en k conjuntos disjuntos de igual tamaño, T_1,..,T_k. Se realizan k experimentos, usando como conjunto de entrenamiento en la iteración i-ésima $\bigcup_{j \neq i} T_j$ y como conjunto de test T_i. Cada algoritmo da lugar a una muestra de k estimaciones del error, y las diferencias entre dos algoritmos se juzgan mediante un contraste acerca de las diferencias entre las medias o las medianas del error muestral, como se verá a continuación.

La mayor ventaja de este diseño experimental es que las estimaciones del error sobre los conjuntos de test son independientes (los conjuntos de test no se solapan). Sin embargo, sí existe un cierto solapamiento en lo que se refiere al conjunto de entrenamiento, ya que cada pareja de conjuntos de entrenamiento comparte una alta fracción de los ejemplos. Por este motivo, este diseño experimental no estudia de forma adecuada la variabilidad inducida por la utilización de distintos ejemplos para el entrenamiento. Adicionalmente, existe un claro desequilibrio entre el número de ejemplos utilizado para test y para train cuando $k > 3$. Esta circunstancia tiene dos efectos: por una parte, los algoritmos cuyo error decrece cuanto mayor sea el número de ejemplos utilizados para el train verán estimada de forma optimista su error producido. Por otra parte, esta estimación del error tendrá una mayor variabilidad [4]. Algunos autores [11]

proponen utilizar una estrategia determinista para realizar las particiones del conjunto de ejemplos, con objeto de que las particiones contengan ejemplos lo más diversos que sea posible, dentro de cada una de ellas y, paralelamente, que las particiones sean similares entre sí. Con esto se consigue eliminar la variabilidad en la estimación del error que se produce en determinados algoritmos (los llamados "inestables" [5]). Adicionalmente, la alternativa propuesta en [11], al ser determinista, permite repetir una experimentación sin necesidad de conocer las particiones del conjunto de ejemplos.

Existen más variaciones de la validación cruzada. La técnica *complete cross validation* [16] utiliza todas las posibles particiones del conjunto de ejemplos con un tamaño dado, lo que mejora la estimación del error de generalización. Como el número de particiones sólo es abordable en problemas de dimensión reducida, es posible seleccionar un número menor de particiones, con la ayuda de diferentes criterios [20][21]. El *leave one out* [17][13] es el caso extremo en que cada conjunto de test contiene un único elemento.

Tests empleados en combinación con la validación cruzada: En condiciones bastante generales, podemos afirmar que el objeto de la comparación de dos algoritmos es decidir si el valor medio de su medida de error sobre la población completa coincide, o es distinto [2].

Si se ha seguido el diseño *multifold cross validation*, se dispone de k estimaciones del error de cada algoritmo, como resultado de evaluarlo sobre cada uno de los conjuntos T_i. Ese conjunto de valores puede considerarse, a su vez, como una muestra de k realizaciones independientes de una variable aleatoria "error muestral", asociada al algoritmo. Bajo este punto de vista, si se desea contrastar que dos algoritmos de aprendizaje son distintos, es válido definir como hipótesis nula del contraste la afirmación "Las dos muestras de errores proceden de poblaciones con medias iguales". Si los datos están apareados (lo que ocurre si los dos algoritmos se han probado sobre las mismas particiones) las dos muestras de errores pueden restarse elemento a elemento, con lo que la hipótesis nula equivalente sería "La diferencia entre los errores muestrales de ambos algoritmos tiene media cero".

Si los errores muestrales de los dos algoritmos siguiesen una distribución normal, el test más potente para contrastar dicha hipótesis, bajo condiciones muy generales, es el test t [9]. Dado que ninguno de los parámetros de la población de errores muestrales es conocido, el número de grados de libertad del estadístico t sólo depende de que las muestras estén apareadas y de que las varianzas de las poblaciones sean iguales o distintas; esto último suele decidirse mediante un test F [27].

Existen numerosos contrastes de bondad de ajuste que pueden aplicarse para decidir si las muestras son normales. Uno de los más utilizados es el de Kolmogorov-Smirnov [6], aunque es conocido que, si la media y la varianza de la población son estimadas a partir de la muestra, como es el caso en este diseño, el test es conservador; la tendencia actual es usar en su lugar el test de Shapiro-Wilk [26] o bien el test omnibus de D'Agostino-Pearson [10].

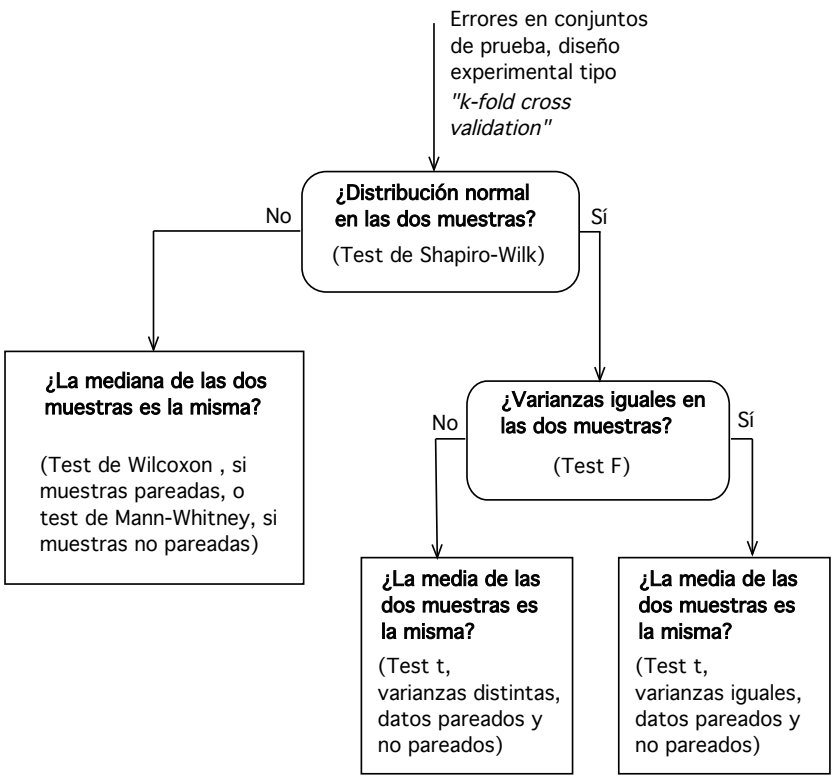

Figura 1. Esquema de los test realizados en el diseño experimental tipo "validación cruzada".

En el caso de que alguna de las muestras no sea normal, el test t no es aplicable, y debe recurrirse a contrastar la hipótesis de que las medianas de las distribuciones del error son iguales, mediante un test no paramétrico. En el caso de que las muestras estén apareadas, puede emplearse un test de signos para la mediana de las diferencias o bien un test Wilcoxon o de rangos signados [33]. Para muestras no apareadas, los tests más frecuentes son el de la mediana y el de Mann-Whitney [19].

Como resumen, en la figura 1 se muestra un esquema con todas las decisiones que se deben tomar cuando se comparan dos algoritmos mediante validación cruzada.

2.2. 5x2cv

En [12] se analizó el comportamiento del método *k-fold cross validation*, combinado con el empleo de un test t. En ese trabajo se puso de manifiesto que, dado que en el numerador del estadístico t aparece la media de las diferencias del error entre los dos algoritmos, y en el denominador la varianza, cuando la estimación

de la varianza era moderadamente baja, una mala estimación de la media provocaba picos en los valores del estadístico t.

Dietterich propuso en ese trabajo sustituir el numerador del estadístico por la diferencia en el error de uno sólo de los experimentos (en lugar de la media de todos ellos) y justificó que es más efectivo realizar $k/2$ ejecuciones de un test *2-fold cross validation*, con diferentes permutaciones de los datos, que realizar un test *k-fold cross validation*. Como solución de compromiso entre la potencia del test y el tiempo de cálculo, propone realizar 5 ejecuciones de un test de validación cruzada con $k = 2$, de ahí el nombre 5x2cv. Los resultados de las 5 permutaciones se combinan mediante el estadístico 5x2cv-t, definido por el mismo autor, que sigue una distribución t con 5 grados de libertad.

Con posterioridad a la definición del diseño experimental 5x2cv, en [1] propuso reemplazar el estadístico 5x2cv-t por una variante que no dependiese del orden en que se realizasen los experimentos. El nuevo estadístico se denominó 5x2cv-f, ya que sigue una distribución $F_{10,5}$. En el mismo estudio se justificó también que el test 5x2cv-f es más potente que el 5x2cv-t bajo ciertas condiciones.

3. Estudio empírico

A continuación se aplicarán los diseños experimentales mencionados en la sección anterior a varios problemas sintéticos, para comprobar empíricamente las propiedades de cada uno. En este estudio, de carácter preliminar, nos hemos limitado a estudiar problemas de clasificación, si bien la misma metodología podría extenderse de forma inmediata a otro tipo de problemas de aprendizaje.

3.1. Descripción de los problemas

En la figura 2 se muestran gráficamente los tres problemas usados en este estudio. El problema al que llameremos "A", definido en [14], consiste en una muestra de tamaño 500 de una población en la que existen dos clases equiprobables, con distribución normal bidimensional, medias $(0, 0)$ y $(2, 0)$, y matrices de covarianza diagonales, de valores I y $4I$, respectivamente. El problema B se ha construido a partir del problema A, desplazando 0.25 unidades a la izquierda a cada punto de la segunda clase. En el problema C, los mismos puntos del problema A se movieron 0.1 unidades a la derecha. Los tres problemas son cuadráticos, y sus superficies discriminantes óptimas son las curvas mostradas en la parte derecha de la figura. En los tres, la solución lineal es subóptima (con un error bayesiano en torno al 20 %), pero numéricamente es muy próxima a la cuadrática (que está en torno al 18 %).

Aunque las diferencias entre los tres problemas son poco perceptibles visualmente, en el problema B las clases están más solapadas, luego la solución cuadrática está más diferenciada de la lineal, y es más fácil que un contraste distinga entre ambas. En el problema C, la situación es la inversa: la diferencia numérica entre las soluciones lineal y cuadrática es menor, y por tanto el problema es más difícil. La relevancia de estas diferencias se comprobará en la siguiente sección.

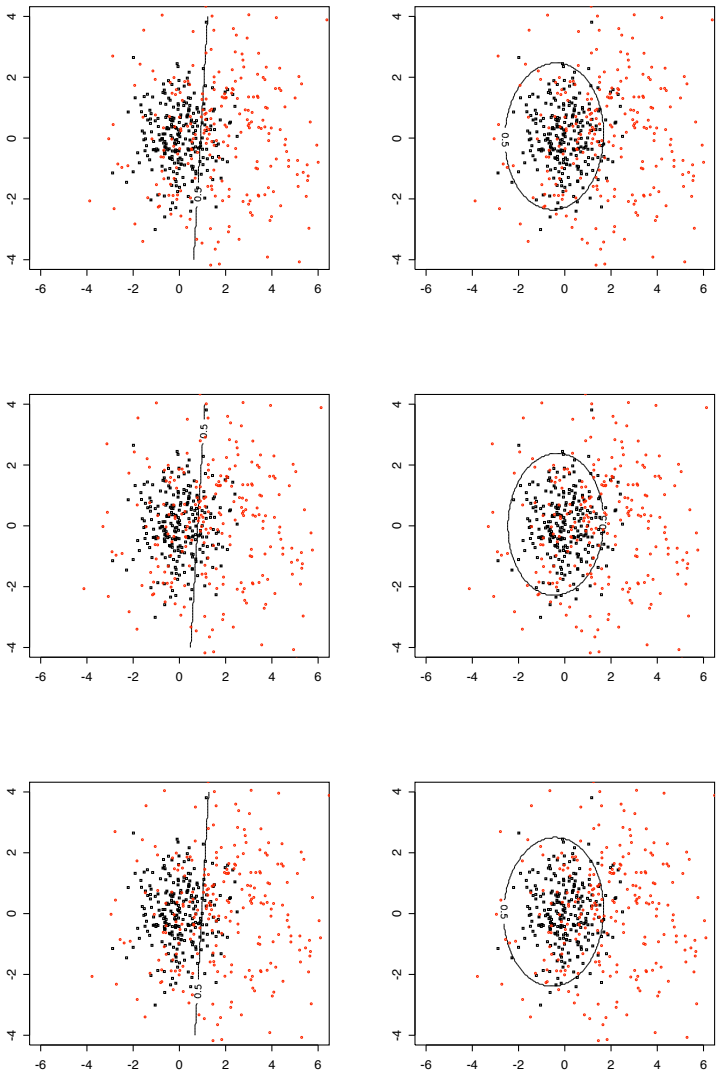

Figura 2. Muestras de datos de los problemas A (arriba), B (centro) y C (abajo) usados en el estudio empírico. Todos los problemas tienen dos clases, con distribución normal y covarianzas I y $4I$. La primera clase está centrada en $(0; 0)$ en los tres problemas, y la segunda en $(2; 0)$, $(1{,}75; 0)$ y $(2{,}1; 0)$ en los problemas A, B y C, respectivamente. El problema más sencillo, desde el punto de vista de los tests estadísticos, es el B, porque las soluciones lineal y cuadrática tiene errores menos parecidos entre sí. El más difícil es el C, en el que se da la situación opuesta. Pese a que los tres problemas son muy parecidos, el problema B es solucionado por todos los tests probados en este estudio, mientras que todos ellos fallan en el C.

3.2. Experimentos realizados

Todos los diseños experimentales realizados tienen como objeto decidir si el algoritmo lineal tiene distinto error que el cuadrático, para los problemas A, B y C mostrados en la figura 2. Como se conoce que los tres problemas son cuadráticos, todos los tests deberían concluir que los algoritmos son diferentes.

Los diseños experimentales comparados son la validación cruzada, con 10, 30, 50 y 100 particiones (*folds*), asociada a los tests mostrados en la figura 1, más el 5x2cv [12], en su versión original y en su versión actual [1]. La validación cruzada se ha realizado tanto con datos apareados como con datos sin aparear. Esta última situación se ha estudiado para comprobar si existe una pérdida de potencia importante cuando los algoritmos se entrenan sobre particiones diferentes.

Cada uno de los 30 diseños resultantes se ha repetido 100 veces, sobre permutaciones aleatorias de la muestra. En la tabla 1 se muestra la fracción de los experimentos en que se ha rechazado la hipótesis nula (medias o medianas iguales) que es falsa en todos los casos. En otras palabras, en la tabla se muestra la fracción de experimentos en que el test fue capaz de concluir que la solución cuadrática es mejor que la lineal, para los niveles de significación 0.01, 0.05 y 0.10.

A la vista de la tabla mencionada, existen diferencias importantes en la potencia de los tests, para los valores correspondientes de los parámetros de la distribución y del nivel de significación de esta simulación. Solamente en el problema B se alcanzan potencias cercanas al 90 %. En el problema C, ninguno de los diseños basados en validación cruzada estándar sobrepasa el 5 % de rechazos con un nivel de significación de 0.95, observándose una ventaja evidente de los tests del tipo 5x2cv.

Por el contrario, en el problema más sencillo (B), el diseño experimental basado en validación cruzada proporciona mejores resultados. Es interesante comprobar que el aumento del número de particiones por encima de 30 no supone mejoras en la potencia, lo que parece reforzar la tesis expuesta en [12] de que es más práctico repetir varias veces la validación cruzada con pocas particiones, sobre diferentes permutaciones de los datos, que efectuar un diseño con un número alto de particiones. Por último, constatar que el empleo de datos no apareados influye notablemente en la capacidad del test para distinguir entre resultados diferentes, como cabía esperar.

4. Conclusiones y trabajo futuro

Como se ha mencionado en la introducción, no hay un acuerdo unánime en lo relativo al diseño experimental en problemas de aprendizaje automático. La opción más difundida consiste en combinar la validación cruzada con un test del tipo t, pero el número de particiones que se debe elegir no está bien definido. Por otra parte, cuando se usan algoritmos de aprendizaje estocásticos, es frecuente repetir varias veces del algoritmo de aprendizaje sobre cada partición, lo que complica la determinación de los grados de libertad del test t correspondiente. Pretendemos abordar ese tipo de estudios en trabajos futuros.

Diseño	**Potencia**								
	Problema A			Problema B			Problema C		
	$\alpha = 0{,}01$	$\alpha = 0{,}05$	$\alpha = 0{,}10$	$\alpha = 0{,}01$	$\alpha = 0{,}05$	$\alpha = 0{,}10$	$\alpha = 0{,}01$	$\alpha = 0{,}05$	$\alpha = 0{,}10$
10-cv, apar.	0.04	0.28	0.58	0.29	0.80	0.96	0.00	0.05	0.10
30-cv, apar.	0.04	0.41	0.76	0.54	0.98	1.00	0.00	0.02	0.16
50-cv, apar.	0.03	0.40	0.71	0.48	1.00	1.00	0.00	0.02	0.13
100-cv, apar.	0.01	0.35	0.87	0.63	1.00	1.00	0.00	0.00	0.08
10-cv, no ap.	0.01	0.12	0.23	0.07	0.62	0.87	0.00	0.02	0.09
30-cv, no ap.	0.00	0.01	0.27	0.06	0.60	0.92	0.00	0.00	0.00
50-cv, no ap.	0.00	0.01	0.26	0.02	0.56	0.94	0.00	0.00	0.00
100-cv, no ap.	0.00	0.02	0.25	0.01	0.63	0.95	0.00	0.00	0.00
5x2cv	0.01	0.21	0.38	0.10	0.44	0.64	0.01	0.08	0.19
5x2cvf	0.06	0.26	0.49	0.26	0.62	0.89	0.00	0.13	0.32

Cuadro 1. Estimaciones numéricas de la potencia de los tests asociados a los diseños experimentales más frecuentes. En todos los casos se ha contrastado si el algoritmo lineal es diferente del cuadrático, para los problemas A, B y C mostrados en la figura 2. Cada uno de los 30 diseños tabulados se ha repetido 100 veces, sobre permutaciones aleatorias de la muestra. Se muestra la fracción de los experimentos en que se ha rechazado la hipótesis nula (medias o medianas iguales) que es falsa en todos los casos. Los valores próximos a 1 indican que la potencia del test es mayor, para los valores correspondientes de los parámetros de la distribución y del nivel de significación.

Algunas de las conclusiones obtenidas en la simulación empírica también son discutibles, y merecen un análisis más profundo. Por citar alguna, los buenos resultados obtenidos para los diseños basados en 10 particiones frente a otros, más costosos en tiempo de cálculo, podrían explicarse a partir de las propiedades de los contrastes de bondad de ajuste: ante muestras pequeñas, es difícil que el test rechace la normalidad de la muestra, por lo que, al seguir el esquema de la figura 1, se utiliza el test t con más frecuencia que el test no paramétrico correspondiente, lo que produce estimaciones optimistas de la potencia en el estudio empírico que nos ocupa. Al igual que en caso anterior, esta afirmación también debería ser contrastada con nuevos experimentos.

5. Agradecimentos

Los autores manifiestan a la Dra. Couso Blanco su agradecimiento por los comentarios realizados acerca de este manuscrito. Este trabajo ha sido financiado por el Ministerio de Ciencia y Tecnología, por los proyectos con códigos TIC2002-04036-C05-01, TIC2002-04036-C05-02 y TIC2002-04036-C05-05.

Referencias

1. Alpaydin E.: Combined 5x2cv-F test for Comparing Supervised Classification Learning Algorithms. Neural Computation 11 (1999) 1885-1892
2. Barr, R. S., Golden, B. L., Kelly, J. P., Resende, M. G. C., Stewart Jr., W. R.: Designing and Reporting on Computational Experiments with Heuristic Methods. Journal of Heuristics, 1 (1995) 9-32
3. Blake, C.L., Merz, C.J.: UCI Repository of machine learning databases. http://www.ics.uci.edu/ mlearn/MLRepository.html. University of California, Department of Information and Computer Science (1998)
4. Bradford J. P.: Brodley C. E.: The effect of Instance-Space Partition on Significance. Machine Learning 42 (2001) 269-286
5. Breiman, L.: Bagging predictors. Machine Learning 24 (1996) 123-140
6. Chakravarti, Laha, and Roy. Handbook of Methods of Applied Statistics, Volume I, John Wiley and Sons, (1967). 392-394.
7. Cochran W. G., Cox G. M.: Experimental Designs. Wiley (1992)
8. Cohen, P. R., Empirical Methods for Artificial Intelligence. MIT Press (1995)
9. Cox, D.R. and Hinkley, D.V. Theoretical statistics. London: Chapman & Hall (1974)
10. D'Agostino, R. B. and Stephens, M. A., eds. Goodness-of-fit Techniques. New York: Dekker (1986)
11. Diamantidis N. A., Karlis D., Giakoumakis E. A.: Unsupervised stratification of cross-validation for accuracy estimation. Artificial Intelligence 116 (2000) 1-16
12. Dietterich, T. G.: Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Computation 10 (7) (1998) 1895-1923
13. Geisser, S: The Predictive Sample Reuse Method with Application. J. Amer. Stat. Ass. 70 (1975) 320-328
14. Haykin, S. Neural Networks, A Comprehensive Foundation.Prentice Hall, 1999
15. Holtr R. C.: Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. Machine Learning 11(1) 1993 63-90

16. Kohavi, R.: A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Proceedings of International Joint Conference on Artificial Intelligence (1995)
17. Lachenbruch, P.A., Mickey M. R.:Estimation of Error Rates in Discriminant Analysis, Technometrics 10 (1968) 1-11
18. Lindman H. R.: Analysis of variance in experimental design.Springer-Verlag (1992)
19. Mann, H. B. and Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Statist. 18, (1947) 50-60.
20. Mitchell, T.: Machine Learning. McGraw Hill (1997)
21. Mullin M., Sukthankar R.: Complete Cross-Validation for Nearest Neighbor Classifiers. Proceedings of the International Conference on Machine Learning (2000)
22. Piater, H. J., Cohen, P. R., Zhang, X., Atighetchi, M.: A Randomized ANOVA Procedure for Comparing Performance Curves. Machine Learning: Proceedings of the Fifteenth International Conference (1998)
23. Ross, S. M.: Introduction to probability and statistics for engineers and scientists. Wiley (1987)
24. Ruiz-Maya, L.: Métodos Estadísticos de Investigación (Introducción al Análisis de la Varianza), Instituto Nacional de Estadística. (1986)
25. Salzberg S. L.: On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach. Data minig and Knowledge Discovery 1 (1997) 317-328
26. Shapiro, S. S. and Wilk, M. B. .An analysis of variance test for normality (complete samples)", Biometrika, 52, 3 and 4, (1965) 591-611
27. Snedecor, G. W., Cochran, W. G.: Statistical Methods. Iowa State University Press, Ames, IA. (1989)
28. Schaffer, C.: A conservation law for generalization performance. In Proceedings of the 1994 International Conference on Machine Learning (1994)
29. Stone, M.: Cross-validatory choice and assesment of statistical predictions. J. Roy. Statist. Soc. 36 (1974) 111-147
30. Stone, M. Cross-validation: A review. Mathematische Operationsforschung Statischen, Serie Statistics, 9 (1978) 127-139
31. Tjen-Sien Lim, Wei-Yin Loh, Yu-Shan Shih: A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms. Machine Learning 40(3) (2000) 203-228
32. Whitley D., Watson J. P., Howe A., Barbulescu L.: Testing, Evaluation and Performance of Optimization and Learning Systems. Keynote Address: Adaptive Computing in Design and Manufacturing (2002)
33. Wilcoxon, F. Individual Comparisons by Ranking Methods. Biometrics 1, (1945) 80-83
34. Wolpert, D.H.: On the Connection Between In-Sample Testing and Generalization Error. Complex Systems 6 (1992) 47-94