Un estudio empírico preliminar acerca del uso de los tests de comparaciones múltiples en aprendizaje automático

José Otero
Dept. de Informática
Universidad de Oviedo
jotero@lsi.uniovi.es

Luciano Sánchez
Dept. de Informática
Universidad de Oviedo
luciano@uniovi.es

Jesús Alcalá
Dept. de C.C.I.A.
Universidad de Granada
jalcala@decsai.ugr.es

Resumen

No son infrecuentes los artículos de aprendizaje automático en los que se comparan varios algoritmos. En este tipo de trabajos no sólo debería meditarse con detenimiento el tipo de diseño experimental a utilizar sino que tendría que usarse algún tipo de test de comparaciones múltiples. Existen ejemplos de este tipo de test en la literatura del tema, pero en general no se utilizan en los artículos de aprendizaje automático. 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 en el campo de los análisis clínicos pero con aplicabilidad en otros dominios. El estudio experimental se instrumenta sobre un conjunto de datos sintéticos, con propiedades teóricas conocidas.

1. Introducción

En una fracción importante de los artículos de aprendizaje automático se comparan al menos dos algoritmos o métodos, generalmente el propuesto en el mismo trabajo con los más relevantes del estado del arte. Según [8] existen distintos factores que hacen necesario emplear algún tipo de test estadístico en este tipo de comparaciones, como por ejemplo la métrica del error, la elección de los conjuntos de entrenamiento y test, y la propia naturaleza del algoritmo, cuando este no es determinista.

El estudio aquí presentado se restringe a experimentos consistentes en resolver una serie de problemas (tomados de entre los usuales en la literatura del tema [2]) usando una implementación de un algoritmo, encaminados a determinar cuál es el que resuelve esos problemas cometiendo el menor error posible.

Se entenderá que el *diseño experimental* viene definido por el conjunto de problemas, medidas realizadas, los detalles de la implementación y, en general, el contexto que acompaña a la realización de los experimentos [4, 22]. No existe todavía un consenso entre los investigadores en aprendizaje de máquina sobre cuales son los tests y diseños experimentales a utilizar [8, 24, 20, 35] y se argumenta que frecuentemente se vulneran una o más de las condiciones que han de cumplirse para la aplicación de determinado test estadístico [5, 24]. Es más, en ocasiones se comparan incluso varios algoritmos simultáneamente [21], lo que según algunos investigadores implica utilizar tests estadísticos específicos [18, 15].

Si bien, como ya se ha dicho, no está generalizado el uso de tests estadísticos en aprendizaje automático, se pueden incorporar casi de forma inmediata los tests utilizados en el campo de los ensayos clínicos [6], ya que su forma es similar a la de los experimentos que se realizan en aprendizaje de máquina.

Por todas estas razones es necesario hacer un repaso a la literatura más relevante sobre tests estadísticos y aplicar estos al campo del aprendizaje de máquina, del modo más ortodoxo posible, es decir verificando el cumplimiento de las condiciones paramétricas necesarias para la aplicación correcta de esos tests, tal y como se propone en [13, 24]. Alternativamente se podría utilizar una técnica basada en remuestreo, de modo que las distribuciones de los estadísticos se derivarían de forma empírica [25, 33].

Una situación que se da con frecuencia en apren-

dizaje de máquina ocurre cuando un investigador desea saber si alguno de los métodos (posiblemente el suyo) es significativamente mejor que el resto de los más relevantes del estado del arte.

Una posible forma de plantear este tipo de comparaciones podría ser la siguiente:

- Si el algoritmo propuesto por el investigador es el que posee menor error medio, entonces:
  - Si las diferencias con los demás algoritmos son significativas, el algoritmo propuesto es el mejor de los analizados.
  - Si las diferencias con el mejor o mejores algoritmos no son significativas, el algoritmo propuesto es equivalente al mejor o mejores algoritmos de los analizados.
- En caso contrario, si el algoritmo propuesto por el investigador no es el que posee menor error medio:
  - Si las diferencias con el mejor o mejores algoritmos no son significativas, el algoritmo propuesto es equivalente al mejor o mejores algoritmos de los analizados.
  - En caso contrario, no puede afirmarse que el algoritmo propuesto por el investigador esté entre los mejores.

Esta secuencia de pasos implica una afirmación compuesta de afirmaciones individuales sobre el resultado de comparar simultáneamente un algoritmo con cada uno de los restantes. Parece evidente que, en algún momento, debe realizarse un test de comparación de medias entre el algoritmo propuesto por el investigador y el resto. En la literatura del tema se conoce a este tipo de test como "test de comparaciones múltiple" [18, 15].

En este trabajo se comentan los tests estadísticos de mayor aplicabilidad en los experimentos de aprendizaje de máquina, organizándose la exposición en tres partes. En la primera, se realiza una taxonomía de los tests de comparaciones múltiples más relevantes. En la segunda se explica cómo aplicar estos tests a las experimentaciones comunes en aprendizaje automático. Finalmente, se realizará un estudio empírico de una selección de estos tests estadísticos utilizando un diseño experimental del tipo *multifold cross validation* [29] sobre un problema sintético, de solución conocida, y se extraerán conclusiones sobre la potencia y error de tipo I de estos tests.

2. Test de comparaciones múltiples, estudio bibliográfico.

Existen textos recientes dedicados enteramente a este tema [18, 15] incluso en combinación con remuestreo [33] o en el marco de determinado soft ware [34] en los cuales se pueden encontrar recopilados la mayor parte de los métodos que se citan en este artículo. También se pueden consultar algunos artículos de "review" sobre este tema como [26, 23, 7, 32]. En el contexto del aprendizaje de máquina se puede consultar [19] y fuera de este (se enmarca en el campo de los análisis clínicos) pero con cierta aplicabilidad se tiene [6]. El investigador que desee confrontar su método con los más representativos del estado del arte, puede plantear varios test de hipótesis simultáneos. Por ejemplo, para el caso de n algoritmos, de la forma:

$$
\begin{gathered}
{}_{0i} : \mu_1 = \mu_i \\
{}_{1i} : \mu_1 \neq \mu_i \\
\in 2
\end{gathered}
$$

en donde el algoritmo con el índice 1 se compara con el resto de los n-1 algoritmos. Es importante destacar que, lo que finalmente pretende al investigador, es poder realizar una afirmación compuesta por los resultados de todos los tests (con una determinada seguridad) como la siguiente: "el algoritmo 1 posee distinto error que 2, igual que el 3,...etc".

Sin embargo, cuando se realiza un test múltiple, se pone de relieve el efecto multiplicativo del error de tipo I: si en cada test la probabilidad de cometer un error de tipo I (rechazar la hipótesis nula cuando es cierta, en nuestro caso encontrar diferencias cuando no existen) es α, la probabilidad de no cometer ninguno en tests es $(1-\alpha)^n$, luego la probabilidad de cometer *al menos* un error de tipo I en el conjunto de tests es $\alpha_t = 1-(1-\alpha)^n$. Si el número de experimentos que se comparan es , en total se tienen $= *(-1)$ 2 tests individuales.

En general, este efecto, que es bien conocido en estadística [18, 15], suele ignorarse en aprendizaje de máquina [24].

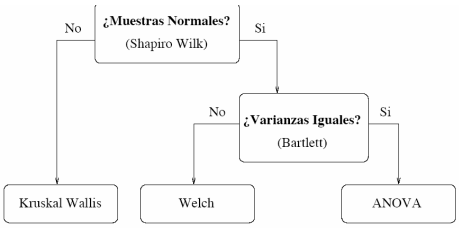

Figura 1: Esquema de los test necesarios para escoger el test F adecuado a las condiciones paramétricas de las muestras.

Existen dos alternativas clásicas, propuestas por Fisher [11] a la hora de abordar este problema, los métodos de dos pasos y los métodos de un paso junto con una tercera clase de métodos, más reciente, denominados multipaso por algunos autores [15] o secuenciales por otros [32].

2.1. Métodos de dos pasos.

Los métodos de dos pasos *protegen* el test de comparaciones múltiples frente al error de tipo I mediante un test F previo (ANOVA, análisis de la varianza, o alguna de sus contrapartidas paramétricas), en el cual se contrasta si existe alguna diferencia entre dos de los algoritmos al menos. Para experimentos las hipótesis nula y alternativa de este test F previo serían:

$$_0 : \mu_1 = \mu_2 = \quad = \mu_r$$
$$_1 : \mu_i \neq \mu_j \qquad\qquad \in 1$$

Es necesario verificar cuales son las condiciones paramétricas que cumplen las muestras, para determinar cual es la secuencia de test que se deben aplicar, como se puede observar en la figura 1. En este trabajo se supone que se cumple la condición de independencia de las muestras.

Una vez realizado alguno de los test anteriores, si se rechaza la hipótesis nula (existen diferencias significativas), se realizan los test de comparaciones múltiples escogiendo el mismo nivel de significación que para el test F. A este tipo de planteamiento corresponde el tests LSD de Fisher [5]. Es importante resaltar que la protección frente a errores de tipo I se produce sólo bajo la hipótesis nula, es decir, frente al hecho de encontrar diferencias entre algún resultado, cuando realmente no existen, en el test F que se aplica en primer lugar. Sin embargo esto no ocurre en el caso de los tests individuales [15, 18], una vez rechazada la hipótesis nula, el test de comparaciones múltiples no está protegido frente al error de tipo I al mismo nivel que el test F, es decir la probabilidad de encontrar diferencias que no existen entre dos algoritmos es mayor que el nivel escogido para la afirmación compuesta. Por lo tanto si la hipótesis nula no es cierta, no se podrá decir de la afirmación a que se hacía mención más arriba en el texto que se realiza con una determinada confianza.

2.2. Métodos de un paso.

Los métodos denominados de un paso tienen en cuenta el efecto multiplicativo del error de tipo I y ajustan los niveles de significación (o los p-valores) de cada test individual con el fin de asegurar que en el test múltiple no se excede determinada probabilidad de ese error. De este modo, si se fija una probabilidad de error de tipo I para el test global de α_t, se tiene para cada test individual $\alpha = 1 - (1 - \alpha_t)^{1/n}$ (siendo el número de comparaciones), lo cual se conoce como el ajuste de Dunn-Sidák [27]. Si se aproxima $(1 - \alpha)^n$ por $1 - \alpha$ se tiene $\alpha = \alpha_t$ lo que se conoce como el ajuste de Bonferroni [3].

2.3. Métodos secuenciales.

Lamentablemente la potencia de un test (1-probabilidad de error de tipo II, aceptar la hipótesis nula cuando es falsa) decrece junto con α, lo cual hace que los mencionados ajustes sean conservadores, es decir, tienden a no rechazar la hipótesis nula [32]. Por este motivo se ha desarrollado el tercer tipo de métodos mencionado al principio de esta sección (métodos secuenciales) que permiten incrementar la potencia de los tests múltiples. Entre estos ajustes se encuentran el de Holm [16], el de Simes-Hochberg [28, 14] o el de Hommel [17]. En esencia se trata de ordenar los p-valores obtenidos en los tests individuales y comenzando por un extremo de la ordenación, aplicar el ajuste de Bonferroni o el de Dunn-Sidak al test actual, rechazando o aceptando la hipótesis correspondiente. A continuación, se

elimina esa hipótesis del test múltiple, actualizando por lo tanto el número de tests (que determina el nivel de significancia de los tests individuales) y se procede con el test múltiple resultante del mismo modo.

Finalmente, en una experimentación de aprendizaje de máquina, no sólo es importante limitar la probabilidad de que se rechace incorrectamente al menos una hipótesis nula cierta, sino que además es importante controlar su número. En este sentido, en [1], se propone una técnica para controlar esta magnitud, que los autores denominan "False Discovery Rate" o FDR por sus siglas. Recientemente en [30] y [10] han aparecido sucesivas modificaciones de esta técnica.

3. Aplicación de los tests de comparaciones múltiples en Aprendizaje Automático

Dado el ámbito de este artículo, es necesario detallar como se aplican los tests comentados cuando se trata de contestar a las preguntas que se formulaban en la sección 1. La necesidad de disponer de una muestra del error medio cometido por cada algoritmo, junto con el pequeño tamaño de los ficheros de datos con los problemas típicos [2], implica la utilización de la validación cruzada como diseño experimental. De esta forma se obtiene para cada algoritmo una muestra del error medio evaluado sobre cada partición de test.

La primera decisión que deberá tomar el investigador es cómo proteger el test múltiple frente a hipótesis nulas rechazadas erróneamente. En función de esta decisión se realizará un test F previo o no.

Si se opta por realizar un test F previo, es necesario verificar las condiciones paramétricas de las muestras, con el fin de elegir el más adecuado según el esquema que se muestra en la figura 1. Si el test F rechaza la hipótesis nula, es decir si los errores medios son significativamente distintos al nivel escogido, se continúa con los test de comparaciones de medias dos a dos, confrontando el algoritmo propuesto con cada uno de los demás, utilizando el mismo nivel de significación. Se utilizará el esquema de la figura 2 propuesto en [13] en esta parte del test. Para cada test de hipótesis se tomarán las muestras del error medio correspondientes a los algoritmos que se desean comparar. Como resultado de este proceso, finalmente se tienen tantos p-valores como tests se han realizado, y en función del nivel de significación elegido se aceptarán o no las hipótesis nulas asociadas. En este momento, se puede comenzar a responder a las preguntas formuladas en la sección 1:

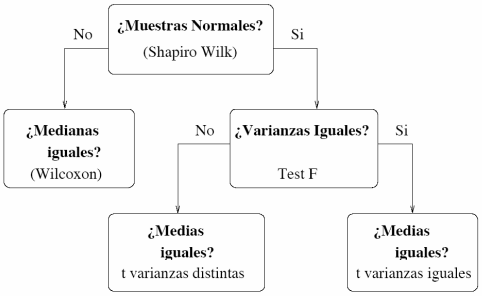

Figura 2: Esquema de los test realizados en el diseño experimental tipo "validación cruzada".

- Si el algoritmo propuesto es el de menor error muestral:
  - Si se rechaza la hipótesis nula del test F y se rechaza la hipótesis nula del test realizado entre el algoritmo propuesto y el mejor, el algoritmo propuesto es el mejor.
  - Si se rechaza la hipótesis nula del test F y no se rechaza la hipótesis nula del test realizado entre el algoritmo propuesto y el mejor, el algoritmo prsopuesto es equivalente al mejor.
  - Si no se rechaza la hipótesis nula del test F, todos los algoritmos son equivalentes.
- Si el algoritmo propuesto no es el de menor error muestral:
  - Si se rechaza la hipótesis nula del test F y se rechaza la hipótesis del test realizado entre el algoritmo propuesto y el mejor, el algoritmo propuesto no es el mejor.
  - Si se rechaza la hipótesis nula del test F y no se rechaza la hipótesis del test realizado entre el algoritmo propuesto y el mejor, el algoritmo propuesto es equivalente al mejor.

- Si no se rechaza la hipótesis nula del test F, todos los algoritmos son equivalentes.

Si se opta por no realizar el test F previo y se prefiere ajustar los p-valores de los test individuales, se utilizará alguno de los métodos de ajuste comentados. En cualquier caso se obtendrá también una serie de p-valores que, determinarán si se rechaza o no cada hipótesis nula en función de su comparación con un determinado nivel de significación (posiblemente distinto para cada hipótesis). Finalmente, si el investigador decide controlar el número de hipótesis nulas erróneamente rechazadas deberá utilizar uno de los métodos que controlan el FDR.

En cualquiera de estos dos últimos casos, también se puede contestar a las preguntas que se hace el investigador normalmente, el esquema similar es el comentado para la prueba F, sin más que omitir esta.

4. Estudio empírico

En la literatura del tema que se ha revisado, no se ha encontrado el uso sistemático de este tipo de tests en el campo del aprendizaje automático. La referencia más destacable en este sentido es [19], pero no se centra en la comparación de algoritmos. Por este motivo se propone en este artículo la realización de un estudio empírico que sirva para valorar su aplicabilidad en este campo. El objeto del estudio empírico es obtener una estimación de la potencia y error de tipo I de algunos de los tests de hipótesis múltiples que se han comentado, en un problema de aprendizaje automático típico. Se pretende investigar el efecto del número de particiones utilizado, del nivel de significación obtenido y si existe diferencia entre las dos familias de tests más destacadas, la formada por los que realizan un test F previo y la formada por los que ajustan los p-valores.

4.1. Experimentos realizados

Se han realizado dos experimentaciones, en cada uno de ellos se han comparado cuatro algoritmos, utilizando dos tests seleccionados de entre las dos familias mencionadas. Cada test se ha repetido 100 veces en cada una de las condiciones en las que se quiere probar. En cada repetición se comprueba si ha distinguido correctamente los algoritmos con distinto error. La fracción de ocasiones en que esto es así es la estimación de la potencia utilizada.

El problema usado en este estudio, definido en [12], 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 y 4 , respectivamente. El problema es cuadrático, 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 %).

Para realizar un estudio completo de la potencia de los tests, habría que contrastar su comportamiento en todas las configuraciones posibles de la *verdad* [9]. En este trabajo, por cuestiones de espacio sólo se contrasta el comportamiento en dos configuraciones, en un caso el algoritmo propuesto no es significativamente diferente (en el error medio cometido) de uno de los restantes, en el otro caso, el algoritmo no es diferente de dos de los restantes. Los algoritmos comparados en el primer caso son los siguientes: cuadrático (entrenado), cuadrático exacto (calculado analíticamente), lineal y k-vecinos. El test múltiple debería contestar que no hay diferencias entre los dos primeros y que si las hay con el tercero y cuarto. En segundo caso, los tres clasificadores con el mismo error sencillamente contestaban de forma aleatoria y equiprobable la clase a que pertenecía cada punto. El cuarto clasificador que se utilizó fue el k-vecinos. Para esta configuración el test debería resolver que no existen diferencias entre los tres primeros pero si entre el primero con el último.

Los tests comparados son los siguientes:

- Método de Holm [16].
- Test F seguido de la batería de tests propuesta en [13]. Es una modificación del test LSD (Least Significant Difference) original de Fisher, en el que se utilizaba un test t.

Cada test se prueba con los niveles 0.01, 0.05 y 0.1 y con 10, 30, 50 y 100 particiones.

4.2. Resultados

Los resultados de la primera experimentación se muestran en la tabla 1 en la que las entradas son

Folds	Nivel		
	0.01	0.05	0.10
10	0.20 0.06	0.70 0.48	0.85 0.68
30	0.10 0.02	0.70 0.36	1.00 0.76
50	0.05 0.00	0.75 0.38	0.95 0.78
100	0.05 0.02	0.90 0.50	1.00 0.94

Cuadro 1: Comparación de las potencias del test protegido (izquierda) y del ajuste de Holm (derecha), primera experimentación.

el número de particiones y el nivel del test. En cada casilla se muestra la potencia estimada para el test de Holm (derecha) y el protegido por la prueba F previa (izquierda). En dicha tabla se aprecia como la potencia aumenta conforme el nivel del test también aumenta, lo cual se cumple en general para cualquier tipo de test. Respecto a la otra entrada de la tabla, el aumento del tamaño de las muestras afecta de forma distinta a un tipo de test y a otro. En general el test protegido no parece ver afectada su potencia por el tamaño de las muestras como sería de esperar (mayor potencia a mayor tamaño), salvo quizás el caso del nivel 0.05 y tamaño muestral 100. En el caso del ajuste de Holm, el efecto sólo se manifiesta en la última columna de la tabla, en donde se reflejan los resultados para el nivel 0.10. El comportamiento observado en los experimentos está completamente de acuerdo con lo expuesto en la literatura del tema [15, 18], al realizar el ajuste de los p-valores (o de forma equivalente, del nivel de significación), el valor obtenido (que es el que se utiliza en la comparación con el nivel de significación) es mayor, de modo que es más probable que no se rechace la hipótesis nula, por lo que la potencia baja. Las mismas fuentes rechazan el empleo de un test F como medida de protección del test global contra el error de tipo I en las situaciones en donde la hipótesis nula global (no existen diferencias) se rechaza.

En el experimento cuyos resultados se muestran en la tabla 1, no se produjo ningún error de tipo I en el único test en donde podría producirse, la comparación del clasificador cuadrático y el cuadrático exacto. Este tipo de error se produce con mayor frecuencia cuando existe un número mayor de hipótesis nulas verdaderas. Por este motivo se planteó la

Folds	Nivel		
	0.01	0.05	0.10
10	0.97 0.98	0.89 0.95	0.79 0.90
30	1.00 1.00	0.97 0.99	0.83 0.97
50	0.95 0.97	0.88 0.92	0.84 0.87
100	0.97 0.99	0.90 0.94	0.82 0.90

Cuadro 2: Comparación de las potencias del test protegido (izquierda) y del ajuste de Holm (derecha), segunda experimentación.

Folds	Nivel		
	0.01	0.05	0.10
10	0.03 0.02	0.11 0.05	0.21 0.10
30	0.00 0.00	0.03 0.01	0.17 0.03
50	0.05 0.03	0.12 0.08	0.16 0.13
100	0.03 0.01	0.10 0.06	0.18 0.10

Cuadro 3: Comparación de los errores de tipo I del test protegido (izquierda) y del ajuste de Holm (derecha), segunda experimentación.

segunda experimentación en la que se comparaban tres clasificadores con el mismo error con otro que no tenía el mismo error. Los resultados son reveladores: como se dijo anteriormente, la potencia varía según la configuración de las hipótesis nulas, según se puede apreciar en la tabla 2 la potencia del test de Holm se ha incrementado hasta igualar e incluso superar al otro test. Además se comprueba también que en este caso el test protegido por una prueba F previa no controla en error de tipo I: como se puede ver en la tabla 3 el error del test protegido por la prueba F previa tiene un error de tipo I del orden del doble de su nivel, mientras que el del test de Holm se mantiene contenido dentro de ese límite.

5. Conclusiones y trabajo futuro

Como se ha mencionado en la introducción, no hay un acuerdo unánime en lo relativo al diseño experimental que debe utilizarse 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. En este trabajo se ha seguido lo expuesto en [13] en lo que se refiere a la necesaria observación de las condiciones paramétricas de aplicabilidad de los distintos tests.

Adicionalmente, cuando se comparan varios algoritmos simultáneamente y se realizan afirmaciones conjuntas sobre el comportamiento de un algoritmo respecto a los restantes, es necesario utilizar algún tipo de test de comparaciones múltiples. La realización de una prueba F previa a las comparaciones múltiples sólo asegura protección frente al error de tipo I si no se rechaza la hipótesis nula de esta. Por lo tanto, se desaconseja esta alternativa y se recomienda el ajuste de p-valores (o de forma equivalente, del nivel de significación) en su lugar. Si se quiere ir más allá en el control de los errores de tipo I individuales, se debe utilizar alguno de los métodos que controla el FDR.

El estudio de los posibles tests estadísticos aplicables a la investigación en aprendizaje de máquina no está completo con el breve repaso que se ha realizado en este trabajo. Existe al menos una familia de tests estadísticos que puede contribuir al análisis de los experimentos realizados en este ámbito, los tests diseñados específicamente para ordenar medias [31]. Por otra parte, en base a los resultados presentados en [25] parece conveniente realizar una experimentación semejante a la que se ha presentado aquí, incorporando los tests basados en remuestreo. Finalmente es posible que en ciertos experimentos sea deseable controlar el número de hipótesis nulas rechazadas incorrectamente, de modo que se debería utilizar el método propuesto en [1].

6. 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 el proyecto con el código TIC2002-04036-C05-05.

Referencias

[1] Benjamini, Y., Hochberg, T. 1995. Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. J. Royal Stat. Soc. B 85: 289300.

[2] Blake, C.L., Merz, C.J. UCI Repository of machine learning databases. http://www.ics.uci.edu. University of California, Department of Information and Computer Science (1998)

[3] Bonferroni C.E. Teoria statistica delle classi e calcolo delle probabilita. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8:3-62. 1936.

[4] Cochran W.G., Cox G.M. Experimental Designs. Wiley (1992)

[5] Cohen, P.R., Empirical Methods for Artificial Intelligence. MIT Press (1995)

[6] Cook R.J., Farewell V.T. Multiplicity Considerations in the Design and Analysis of Clinical Trials. J. R. Statist. Soc. A. (1996) 159, Part 1, pp. 93-110.

[7] Curran-Everett D. Multiple comparisons: philosophies and illustrations. Am J Physiol Regul Integr Comp Physiol 279: R1-R8, 2000.

[8] Dietterich, T.G. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Computation 10 (7) (1998) 1895-1923

[9] Einot, I. and Gabriel, K. R. (1975). A Study of the Powers of Several Methods of Multiple Comparison, Journal of the American Statistical Association, 70, page 351.

[10] Fernando, R.L., D. Nettleton, B.R. Southey, J.C.M. Dekkers, M.F. Rothschild, M. Soller. 2004. Controlling the proportion of false positives in multiple dependent tests. Genetics 166:611-619.

[11] Fisher, R. A. (1971). The design of experiments (9th ed.). New York, Hafner Publishing Company.

[12] Haykin, S. Neural Networks, A Comprehensive Foundation.Prentice Hall, 1999

[13] Herrera F., Hervás C., Otero J., Sánchez L.. Un estudio empírico preliminar sobre los tests estadísticos más habituales en el aprendizaje automático. Tendencias de la Minería de Datos en España 403-412. Lecture Notes in Computer Science. ed. Digital @3D.

[14] Hochberg, Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika, 75: 800-803. 1988.

[15] Hochberg, Y., Tamhane A. C., Multiple comparison procedures, 1987,John Wiley & Sons, Inc.

[16] Holm S. A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics, 6, 65-70. 1979.

[17] Hommel, G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika, 75, 383-386. 1988

[18] Hsu, J.C. (1996). Multiple Comparisons: Theory and Methods, Chapman and Hall,London.

[19] Jensen D., Cohen P. R. (2000). Multiple Comparisons in Induction Algorithms. Machine Learning 38: 309-338.

[20] Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Proceedings of International Joint Conference on Artificial Intelligence (1995)

[21] 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

[22] Lindman H.R. Analysis of variance in experimental design.Springer-Verlag (1992)

[23] Rafter J.A., Abell M. L., James P. Braselton. Multiple Comparison Methods for Means. SIAM Review Volume 44, Number 2 pp. 259-278. 2002 Society for Industrial and Applied Mathematics

[24] Salzberg S. L. On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach. Data minig and Knowledge Discovery 1 (1997) 317-328

[25] Sánchez L., Otero J., Alcalá J. Assesing the differences in accuracy between GFSs with bootstrap tests. Actas del congreso Eusflat 2005.

[26] Sarkar S.K., Temple University, Philadelphia, Pennsylvania, U.S.A. Research Report.

[27] Sidak Z, Rectangular Confidence Regions for the Means of Multivariate Normal Distributions, American Statistical Association, 62, 626-633. 1967.

[28] Simes RJ. An improved Bonferroni procedure for multiple tests of significance. Biometrika, 73, 751-754. 1986.

[29] Stone, M. Cross-validation: A review. Mathematische Operationsforschung Statischen, Serie Statistics, 9 (1978) 127-139

[30] Storey J.D. 2002. A direct approach to false discovery rates. Journal of the Royal Statistical Society Series B: 64: 479498.

[31] Swisher J.R., Jacobson S.H., A survey of ranking, selection, and multiple comparison procedures for discrete-event simulation, Proceedings of the 31st conference on Winter simulation: Simulation—a bridge to the future, p.492-501, December 05-08, 1999, Phoenix, Arizona, United States.

[32] Walsh B. Multiple Comparisons: Bonferroni Corrections and False Discovery Rates. Lecture Notes for EEB581. Department of Ecology and Evolutionary Biology. University of Arizona. May 2004.

[33] Westfall P.H., Young S.S. (1993) Resampling-based Multiple Testing. Wiley, New York.

[34] Westfall, P.H., Tobias, R.D., Rom, D., Wolfinger, R.D. und IIochberg, Y. (1999) Multiple comparisons and multiple tests using the SAS system. Cary, NC.

[35] 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)