Un tutorial sobre el uso de test estadísticos no paramétricos en comparaciones múltiples de metaheurísticas y algoritmos evolutivos

Joaquín Derrac, Salvador García, Daniel Molina, Francisco Herrera

Resumen— **Recientemente, los test estadísticos no paramétricos han emergido como una metodología eficaz, robusta y asequible para la evaluación de nuevas propuestas de metaheurísticas y algoritmos evolutivos, alcanzando gran popularidad en la literatura. En este trabajo se revisan los métodos no paramétricos de comparaciones múltiples más representativos, aplicados a un caso de estudio. La revisión concluye con una compilación de preguntas frecuentes, completando la utilidad del tutorial como guía para investigadores que deseen contrastar los resultados de sus estudios experimentales de forma rigurosa.**

Palabras clave— **Comparaciones múltiples, Tests no paramétricos, Tests estadísticos, Metaheurísticas, Algoritmos Evolutivos**

I. Introducción

El diseño de experimentos es un tarea de importancia capital en el desarrollo de nuevos propuestas. La validación de nuevos algoritmos frecuentemente requiere la definición de un marco experimental exhaustivo, incluyendo un amplio abanico de problemas y algoritmos del estado del arte. La parte crítica de estas comparaciones recae en la validación estadística de los resultados, contrastando las diferencias encontradas entre métodos.

En este sentido, es importante contar con una metodología estadística robusta que avale las conclusiones obtenidas. Dentro del elenco de técnicas disponibles, destacan los tests no paramétricos [12] debido a su flexibilidad y a las pocas restricciones de uso que presentan (en contraste a su contrapartida paramétrica, la cual sufre a menudo problemas derivados de la imposibilidad de cumplir las propiedades de independencia, normalidad y homocedasticidad, necesarias para su uso [9]).

Las primeras apariciones de metodologías estadísticas se limitan a realizar comparaciones del algoritmos por pares. El Test de Signos y el Test de Wilcoxon (y su contrapartida paramétrica, el t-test) permiten evaluar el rendimiento de dos algoritmos en un entorno multiproblema. Un número razonablemente elevado de problemas permite a estos tests establecer una comparación fiable entre pares de técnicas, produciendo un p-valor para reflejar el resultado de dicha comparación. Sin embargo, es habitual encontrar situaciones en que las comparaciones por pares son insuficientes. Por ejemplo, la presentación de una nueva propuesta generalmente requerirá la realización de una comparación simultánea contra varios métodos del estado del arte. En este tipo de casos, las comparaciones por pares suelen ser insuficientes, ya que los p-valores obtenidos al realizar cada comparación no pueden acumularse de forma rigurosa. El error cometido por considerar simultáneamente una familia de hipótesis relacionadas en lugar de tomarlas individualmente hace deseable la utilización de técnicas de comparación múltiple para obtener un análisis más preciso.

En este trabajo se revisan un conjunto de tests de comparaciones múltiples, aplicados sobre un estudio experimental común (Sección II). El cuerpo principal (Sección III) lo componen el Test de Friedman y sus dos versiones avanzadas: El Test de Friedman Alineado, y el Test de Quade. La Sección IV presenta varios procedimientos posteriores de identificación de diferencias (*post-hoc*). Seguidamente, la Sección V incluye otros de comparaciones múltiples aplicables en éste contexto. En la Sección VI se revisan algunas preguntas frecuentes sobre el uso de esta metodología, de especial relevancia. Finalmente, la Sección VII concluye el trabajo.

II. Marco experimental

Para ilustrar el funcionamiento de los tests, se ha seleccionado como caso de uso una comparación basada en los 25 problemas presentados en la Sesión Especial de Optimización de Parámetros Reales del Congreso IEEE sobre Computación Evolutiva de 2005 (CEC'2005). En la comparación, se han empleado 4 algoritmos clásicos: Un algoritmo de Optimización de Nube de Partículas (**PSO**), un algoritmo Genético Estacionario (**SSGA**), un algoritmo de Búsqueda Dispersa con operador de cruce BLX (**SS-BLX**) y un modelo de Evolución Diferencial con operador de cruce *Rand/1/exp* (**DE-EXP**).

Como medida de rendimiento, se ha recogido para cada problema el error medio obtenido en 50 ejecuciones de cada técnica. Se han empleado las versiones de 10 dimensiones de los problemas, y las ejecuciones han finalizado al alcanzar un error inferior a $10^{-}8$ o

J. Derrac y F. Herrera son miembros del Dpto. de Ciencias de la Computación e Inteligencia Artificial, CITIC-UGR. Universidad de Granada. 18071, Granada, España. Email:{jderrac,herrera}@decsai.ugr.es.

S. García es miembro del Dpto. de Informática, Universidad de Jaén, 23071, Jaén, España. Email: sglopez@ujaen.es

D. Molina es miembro del Dpto. de Lenguajes y Sistemas Informáticos, Universidad de Cádiz, 11003, Cádiz, España. Email: daniel.molina@uca.es

tras consumir 100000 evaluaciones [1]. Estos resultados se mostrarán en las diferentes secciones del trabajo, según vayan siendo empleados por los distintos procedimientos estadísticos.

III. Tests de comparaciones multiples

En esta sección se presenta el Test de Friedman como procedimiento para realizar comparaciones múltiples entre diferentes algoritmos. Una vez revisado, se describirán los Tests de Friedman Alineado y Quade, como versiones avanzadas del mismo.

A. Test de Friedman

El Test de Friedman [7], [8] trabaja asignando rankings r_{ij} a los resultados obtenidos por cada algoritmo j en cada problema i. Esto es, para cada problema, se asigna un ranking $1 \leq r_{ij} \leq k$, donde k es el número de algoritmos a comparar. Estos rankings se asignan de forma ascendente, es decir, 1 al mejor resultado, 2 al segundo, etc. (en caso de haber empates, se asignan rankings medios).

El Test de Friedman requiere el cálculo de los rankings medios de los algoritmos sobre los n problemas,

$$R_j = \frac{\sum_{i=1}^{n} r_{ij}}{n} \tag{1}$$

La hipótesis nula que indica que todos los algoritmos se comportan similarmente, por lo que sus rankings R_j deben ser similares. Siguiendo esta hipótesis, el estadístico de Friedman

$$F_F = \frac{12n}{k(k+1)} \left[\Sigma_j R_j^2 - \frac{k(k+1)^2}{4} \right] \tag{2}$$

se distribuye de acuerdo a una distribución χ^2 con $k-1$ grados de libertad. Este estadístico fué mejorado a su vez por Iman y Davenport, quienes mostraron que el estadístico de Friedman presenta un comportamiento demasiado conservativo. Para evitar éste problema, propusieron otro estadístico más ajustado

$$F_{ID} = \frac{(n-1)F_F}{n(k-1) - F_F} \tag{3}$$

que se distribuye de acuerdo a una distribución F con $k-1$ y $(k-1)(n-1)$ grados de libertad. La Tabla A10 de [14] permite consultar los valores críticos de este test. Si se rechaza la hipótesis nula, se puede proceder con un test *post-hoc* para encontrar diferencias a posteriori (esto se detallará más adelante, en la Sección IV).

La Tabla I muestra los rankings obtenidos para el caso de estudio. Los rankings medios proporcionan una comparación interesante de los algoritmos. En media, **DE-EXP** obtuvo el mejor ranking (1,72), seguido por **SS-BLX** y **SSGA** (que muestran un rendimiento más parecido), mientras que **PSO** ofrece el peor resultado. En este punto, el Test de Friedman procede comprobando si los rankings medios obtenidos son significativamente diferentes del ranking medio esperado bajo la hipótesis nula, $R_j = 2,5$

TABLA I
Rankings del Test de Friedman.

Algoritmo	**PSO**	**SSGA**	**SS-BLX**	**DE-EXP**
F1	1,23E-01 (3)	8,42E-06 (2)	3,40E+02 (4)	8,26E-06 (1)
F2	2,60E+01 (3)	8,72E-02 (2)	1,73E+03 (4)	8,18E-06 (1)
F3	5,17E+07 (2)	7,95E+07 (3)	1,84E+08 (4)	9,94E+04 (1)
F4	2,49E+03 (3)	2,59E+00 (2)	6,23E+03 (4)	8,35E-06 (1)
F5	4,10E+05 (4)	1,34E+05 (3)	2,19E+03 (2)	8,51E-06 (1)
F6	7,31E+05 (4)	6,17E+03 (2)	1,15E+05 (3)	8,39E-06 (1)
F7	2,68E+02 (1)	1,27E+06 (2,5)	1,97E+06 (4)	1,27E+06 (2,5)
F8	2,04E+04 (2,5)	2,04E+04 (2,5)	2,04E+04 (2,5)	2,04E+04 (2,5)
F9	1,44E+04 (4)	7,29E-06 (1)	4,20E+03 (3)	8,15E-06 (2)
F10	1,40E+04 (3)	1,71E+04 (4)	1,24E+04 (2)	1,12E+04 (1)
F11	5,59E+03 (4)	3,26E+03 (3)	2,93E+03 (2)	2,07E+03 (1)
F12	6,36E+05 (4)	2,79E+05 (3)	1,51E+05 (2)	6,31E+04 (1)
F13	1,50E+03 (4)	6,71E+02 (3)	3,25E+02 (1)	6,40E+02 (2)
F14	3,30E+03 (4)	2,26E+03 (1)	2,80E+03 (2)	3,16E+03 (3)
F15	3,40E+05 (4)	2,92E+05 (2)	1,14E+05 (1)	2,94E+05 (3)
F16	1,33E+05 (4)	1,05E+05 (2)	1,04E+05 (1)	1,13E+05 (3)
F17	1,50E+05 (4)	1,19E+05 (2)	1,18E+05 (1)	1,31E+05 (3)
F18	8,51E+05 (4)	8,06E+05 (3)	7,67E+05 (2)	4,48E+05 (1)
F19	8,50E+05 (3)	8,90E+05 (4)	7,56E+05 (2)	4,34E+05 (1)
F20	8,51E+05 (3)	8,89E+05 (4)	7,46E+05 (2)	4,19E+05 (1)
F21	9,14E+05 (4)	8,52E+05 (3)	4,85E+05 (1)	5,42E+05 (2)
F22	8,07E+05 (4)	7,52E+05 (2)	6,83E+05 (1)	7,72E+05 (3)
F23	1,03E+06 (4)	1,00E+06 (3)	5,74E+05 (1)	5,82E+05 (2)
F24	4,12E+05 (4)	2,36E+05 (2)	2,51E+05 (3)	2,02E+05 (1)
F25	5,10E+05 (1)	1,75E+06 (3)	1,79E+06 (4)	1,74E+06 (2)
R_j	3,38	2,56	2,34	1,72

$$F_F = \frac{12 \cdot 25}{4(4+1)} \cdot \left[(3,38)^2 + (2,56)^2 + (2,34)^2 + (1,72)^2 - \frac{4(4+1)^2}{4} \right] = 21,18$$

$$F_{ID} = \frac{(25-1)21,18}{25(4-1)-21,18} = 9,44$$

Con cuatro algoritmos y 25 conjuntos, F_{ID} se distribuye de acuerdo a una distribución F con $4-1=3$ y $(4-1)(25-1)=72$ grados de libertad. El p-valor calculado usando la distribución $F(3,72)$ es 2,46E-05, por lo que la hipótesis nula se rechaza con una alta probabilidad.

B. Test de Friedman Alineado

El Test de Friedman está orientado a realizar comparaciones intra-conjunto (comparaciones entre rendimientos de algoritmos en un solo conjunto), sin considerar las interrelaciones que puedan existir entre los conjuntos de la prueba completa. Cuando el número de algoritmos en la comparación es pequeño (3,5,...), este procedimiento tiene cierta desventaja. En estos casos, en los que la comparación entre conjuntos de datos es más deseable, es recomendable emplear el Test de Friedman Alineado [11].

En esta técnica, se calcula el rendimiento medio alcanzado por cada algoritmo en cada problema (valor de localización). Después, se calculan las diferencias entre el rendimiento obtenido por cada algoritmo con respecto al valor de localización. Este paso se repite para todos los algoritmos y problemas. Las

[1] Una descripción más detallada del marco experimental empleado, incluyendo configuraciones y parámetros de los algoritmos, puede encontrarse en [5]

TABLA II
RANKINGS DEL TEST DE FRIEDMAN ALINEADO.

Algoritmo	**PSO**	**SSGA**	**SS-BLX**	**DE-EXP**
F1	-8,50E+01(51)	-8,51E+01(50)	2,55E+02(57)	-8,51E+01(49)
F2	-4,13E+02(45)	-4,39E+02(44)	1,29E+03(63)	-4,39E+02(43)
F3	-2,72E+07(2)	5,50E+05(98)	1,05E+08(100)	-7,88E+07(1)
F4	3,08E+02(59)	-2,18E+03(34)	4,05E+03(67)	-2,18E+03(33)
F5	2,73E+05(92)	-2,20E+03(31)	-1,34E+05(17)	-1,36E+05(16)
F6	5,18E+05(97)	-2,07E+05(13)	-9,84E+04(19)	-2,13E+05(12)
F7	-1,13E+06(3)	1,45E+05(85)	8,40E+05(99)	1,39E+05(84)
F8	4,75E+01(56)	-1,25E+01(54)	-3,25E+01(53)	-2,50E+00(55)
F9	9,74E+03(68)	-4,64E+03(28)	-4,49E+02(42)	-4,64E+03(29)
F10	3,58E+02(60)	3,44E+03(66)	-1,29E+03(37)	-2,50E+03(32)
F11	2,13E+03(65)	-2,05E+02(46)	-5,31E+02(40)	-1,39E+03(36)
F12	3,54E+05(96)	-2,92E+03(30)	-1,32E+05(18)	-2,19E+05(9)
F13	7,18E+02(62)	-1,13E+02(48)	-4,60E+02(41)	-1,44E+02(47)
F14	4,24E+02(61)	-6,17E+02(39)	-8,45E+01(52)	2,78E+02(58)
F15	8,00E+04(78)	3,22E+04(74)	-1,46E+05(15)	3,42E+04(75)
F16	1,95E+04(70)	-8,50E+03(27)	-9,70E+03(26)	-1,30E+03(38)
F17	2,03E+04(72)	-1,09E+04(25)	-1,11E+04(24)	1,78E+03(64)
F18	1,33E+05(82)	8,82E+04(79)	4,87E+04(76)	-2,70E+05(7)
F19	1,17E+05(80)	1,58E+05(87)	2,32E+04(73)	-2,98E+05(6)
F20	1,25E+05(81)	1,63E+05(88)	2,00E+04(71)	-3,08E+05(5)
F21	2,16E+05(90)	1,54E+05(86)	-2,13E+05(11)	-1,56E+05(14)
F22	5,37E+04(77)	-1,55E+03(35)	-7,07E+04(21)	1,86E+04(69)
F23	2,31E+05(91)	2,07E+05(89)	-2,23E+05(8)	-2,15E+05(10)
F24	1,37E+05(83)	-3,93E+04(22)	-2,40E+04(23)	-7,33E+04(20)
F25	-9,38E+05(4)	2,99E+05(94)	3,46E+05(95)	2,94E+05(93)
Suma	1625	1372	1148	905
Media	65	54,88	45,92	36,2

diferencias resultantes (observaciones alineadas) se ordenan desde 1 hasta $k \cdot n$ de forma relativa unas con otras. A partir de ahí, el esquema de ranking es el mismo que el empleado por un procedimiento de comparaciones múltiples con muestras independientes, como el test de Kruskal-Wallis. De este modo, los rankings asignados a las observaciones alineadas se denominan rankings alineados.

El estadístico del Test de Friedman Alineado se define como

$$F_{AR} = \frac{(k-1)\left[\sum_{j=1}^{k} \hat{R}_j^2 - (kn^2/4)(kn+1)^2\right]}{\{[kn(kn+1)(2kn+1)]/6\} - (1/k)\sum_{i=1}^{n} \hat{R}_i^2} \quad (4)$$

donde $\hat{R}_i$ es igual al ranking total del i-ésimo problema y $\hat{R}_j$ es el ranking total del j-ésimo algoritmo.

El estadístico F_{AR} se ajusta a una distribución χ^2 con $k-1$ grados de libertad. Los valores críticos pueden encontrarse en la Tabla A3 de [14]. Si se rechaza la hipótesis nula, se puede proceder con un test *post-hoc* para encontrar diferencias a posteriori (esto se detallará más adelante, en la Sección IV).

A continuación se muestra la aplicación del Test de Friedman Alineado a nuestro caso de estudio. La Tabla II muestra las observaciones alineadas y los rankings alineados (entre paréntesis) obtenidos.

De nuevo, los rankings medios proporcionan una buena comparación de los algoritmos. En media, **DE-EXP** es el mejor con ranking 36,2; **SS-BLX** y **SSGA** muestran un rendimiento similar con rankings 45,92 y 54,88 , mientras que **PSO** aparece en último lugar con ranking 65. Ahora el Test de Friedman Alineado comprueba si la suma de rankings alineados es significativamente diferente al ranking alineado total ${}^{R}j = 1262,5$ esperado bajo la hipótesis nula:

$$\sum_{j=1}^{k} \hat{R}_{\cdot j}^2 = 1625^2 + 1372^2 + 1148^2 + 905^2 = 6659938$$

$$\sum_{i=1}^{n} \hat{R}_{i\cdot}^2 = 207^2 + 195^2 + \ldots + 286^2 = 1054198$$

$$F_{AR} = \frac{(4-1)\left[6659938 - (4 \cdot 25^2/4)(4 \cdot 25 + 1)^2\right]}{\{[4 \cdot 25(4 \cdot 25+1)(2 \cdot 4 \cdot 25+1)]/6\} - (1/4) \cdot 1054198} = 19{,}743$$

Con 4 algoritmos y 24 problemas, F_{AR} se distribuye de acuerdo a una distribución χ^2 con $4-1=3$ grados de libertad. El p-valor calculado usando la distribución $\chi^2(3)$ es 1,92E-04, por lo que la hipótesis nula es rechazada con un alto nivel de significancia.

C. Test de Quade

El Test de Friedman considera que todos los problemas son iguales en importancia. Una alternativa a esto podría tener en cuenta que algunos problemas son más difíciles o que las diferencias registradas en la ejecución de varios algoritmos sobre ellos son más distantes. Así, los rankings calculados en cada problema podrían escalarse dependiendo de las diferencias observadas en los rendimientos de los algoritmos.

El test de Quade [13] lleva a cabo un análisis con rankings ponderados de las muestras de resultados. El procedimiento comienza encontrando los rankings r_{ij} de la misma forma que el Test de Friedman. El siguiente paso requiere los valores originales de rendimiento de los algoritmos x_{ij}. Los rankings se asignan a los problemas de acuerdo al tamaño del rango de la muestra en cada uno. El rango de la muestra en un problema i es la diferencia entre la observación más alta y la más baja en dicho problema

$$\text{Rango en el problema i: } = \max_j x_{ij} - \min_j x_{ij} \quad (5)$$

Obviamente, hay n rangos muestrales, uno por cada problema. Asignamos el ranking 1 al conjunto con el menor rango, el 2 al segundo con menor rango, etc ..., hasta el mayor rango que obtiene un ranking n. Se utilizan rankings medios en caso de empate. Sean $Q_1, Q_2, \ldots, Q_n$ los rankings asignados a los problemas 1, 2, ..., n, respectivamente.

Finalmente, el ranking Q_i se multiplica por la diferencia entre rankings dentro de cada problema i, r_{ij} , y el ranking medio de cada conjunto, $(k+1)/2$ para obtener el producto S_{ij} , donde

$$S_{ij} = Q_i \left[r_{ij} - \frac{k+1}{2} \right] \quad (6)$$

TABLA III
RANKINGS DEL TEST DE QUADE.

Problema	Rango	Q_i	**PSO**	**SSGA**	**SS-BLX**	**DE-EXP**
F1	3,40E+02	2	1,23E-01(1)(6)	8,42E-06(-1)(4)	3,40E+02(3)(8)	8,26E-06(-3)(2)
F2	1,73E+03	5	2,60E+01(2,5)(15)	8,72E-02(-2,5)(10)	1,73E+03(7,5)(20)	8,18E-06(-7,5)(5)
F3	1,84E+08	25	5,17E+07(-12,5)(50)	7,95E+07(12,5)(75)	1,84E+08(37,5)(100)	9,94E+04(-37,5)(25)
F4	6,23E+03	8	2,49E+03(4)(24)	2,59E+00(-4)(16)	6,23E+03(12)(32)	8,35E-06(-12)(8)
F5	4,09E+05	16	4,10E+05(24)(64)	1,34E+05(8)(48)	2,19E+03(-8)(32)	8,51E-06(-24)(16)
F6	7,31E+05	22	7,31E+05(33)(88)	6,17E+03(-11)(44)	1,15E+05(11)(66)	8,39E-06(-33)(22)
F7	1,97E+06	24	2,68E+02(-36)(24)	1,27E+06(0)(60)	1,97E+06(36)(96)	1,27E+06(0)(60)
F8	8,00E+01	1	2,04E+04(0)(2,5)	2,04E+04(0)(2,5)	2,04E+04(0)(2,5)	2,04E+04(0)(2,5)
F9	1,44E+04	9	1,44E+04(13,5)(36)	7,29E-06(-13,5)(9)	4,20E+03(4,5)(27)	8,15E-06(-4,5)(18)
F10	5,94E+03	7	1,40E+04(3,5)(21)	1,71E+04(10,5)(28)	1,24E+04(-3,5)(14)	1,12E+04(-10,5)(7)
F11	3,52E+03	6	5,59E+03(9)(24)	3,26E+03(3)(18)	2,93E+03(-3)(12)	2,07E+03(-9)(6)
F12	5,73E+05	21	6,36E+05(31,5)(84)	2,79E+05(10,5)(63)	1,51E+05(-10,5)(42)	6,31E+04(-31,5)(21)
F13	1,18E+03	4	1,50E+03(6)(16)	6,71E+02(2)(12)	3,25E+02(-6)(4)	6,40E+02(-2)(8)
F14	1,04E+03	3	3,30E+03(4,5)(12)	2,26E+03(-4,5)(3)	2,80E+03(-1,5)(6)	3,16E+03(1,5)(9)
F15	2,26E+05	14	3,40E+05(21)(56)	2,92E+05(-7)(28)	1,14E+05(-21)(14)	2,94E+05(7)(42)
F16	2,92E+04	10	1,33E+05(15)(40)	1,05E+05(-5)(20)	1,04E+05(-15)(10)	1,13E+05(5)(30)
F17	3,14E+04	11	1,50E+05(16,5)(44)	1,19E+05(-5,5)(22)	1,18E+05(-16,5)(11)	1,31E+05(5,5)(33)
F18	4,03E+05	15	8,51E+05(22,5)(60)	8,06E+05(7,5)(45)	7,67E+05(-7,5)(30)	4,48E+05(-22,5)(15)
F19	4,56E+05	19	8,50E+05(9,5)(57)	8,90E+05(28,5)(76)	7,56E+05(-9,5)(38)	4,34E+05(-28,5)(19)
F20	4,71E+05	20	8,51E+05(10)(60)	8,89E+05(30)(80)	7,46E+05(-10)(40)	4,19E+05(-30)(20)
F21	4,29E+05	17	9,14E+05(25,5)(68)	8,52E+05(8,5)(51)	4,85E+05(-25,5)(17)	5,42E+05(-8,5)(34)
F22	1,24E+05	12	8,07E+05(18)(48)	7,52E+05(-6)(24)	6,83E+05(-18)(12)	7,72E+05(6)(36)
F23	4,54E+05	18	1,03E+06(27)(72)	1,00E+06(9)(54)	5,74E+05(-27)(18)	5,82E+05(-9)(36)
F24	2,10E+05	13	4,12E+05(19,5)(52)	2,36E+05(-6,5)(26)	2,51E+05(6,5)(39)	2,02E+05(-19,5)(13)
F25	1,28E+06	23	5,10E+05(-34,5)(23)	1,75E+06(11,5)(69)	1,79E+06(34,5)(92)	1,74E+06(-11,5)(46)
Suma S_j			234	75	-30	-279
Rankings medios $T_j = \frac{W_j}{n(n+1)/2}$			3,22	2,73	2,41	1,64

es un estadístico que representa el tamaño relativo de cada observación dentro de cada problema, ajustado para reflejar la significancia relativa del problema en el que aparece.

Para relacionarlo con Friedman, usaremos el ranking sin ajuste medio:

$$W_{ij} = Q_i\,[r_{ij}] \tag{7}$$

S_j denota la suma para cada clasificador,$S_j = \sum_{i=1}^{n} S_{ij}$ y $W_j = \sum_{i=1}^{n} W_{ij}$, para $j = 1, 2, \ldots, k$. Después, debemos calcular los términos

$$A = n(n+1)(2n+1)k(k+1)(k-1)/72 \tag{8}$$

$$B = \frac{1}{n}\sum_{j=1}^{k} S_j^2 \tag{9}$$

El estadístico es

$$F_Q = \frac{(n-1)B}{A-B} \tag{10}$$

que está distribuido de acuerdo a una distribución F con $k-1$ y $(k-1)(n-1)$ grados de libertad. La tabla de valores críticos para esta distribución puede consultarse en [14], Tabla A10. Si la hipótesis nula es rechazada, se puede proceder con un test *post-hoc* para encontrar diferencias a posteriori (esto se detallará más adelante, en la Sección IV).

La Tabla III muestra un ejemplo del uso del test de Quade sobre el marco experimental descrito. Los rankings medios $T_j = \frac{W_j}{n(n+1)/2}$ pueden ser comparados con los rankings obtenidos por el Test de Friedman clásico. En este caso, **DE-EXP** es el mejor algoritmo con ranking 1,64, **SS-BLX** y **SSGA** obtienen la segunda y tercera posición, con rankings 2,41 y 2,63, respectivamente; y el peor algoritmo es **PSO** con ranking 3,22. El test de Quade comprueba si las suma de rankings ponderados S_j son significativamente diferentes de 0 (hipótesis nula)

$$A = 25(25+1)(2\cdot 25+1)4(4+1)(4-1)/72 = 27625$$

$$B = \frac{1}{25}\left[(234)^2 + (75)^2 + (-30)^2 + (-279)^2\right] = 5564,88$$

$$F_Q = \frac{(25-1)5564,88}{27625 - 5564,88} = 13,166427$$

Con 4 algoritmos y 24 problemas, F_Q se distribuye de acuerdo a una distribución F con $4-1=3$ y $4-1\cdot 25-1=72$ grados de libertad. El p-valor calculado usando la distribución F(3,69) es 6,06E-07, por lo que la hipótesis nula se rechaza con un alto nivel de significancia.

IV. PROCEDIMIENTOS POST-HOC DE IDENTIFICACIÓN DE DIFERENCIAS

Tanto el Test de Friedman como sus dos propuestas avanzadas (Friedman Alineado y Quade) no identifican las diferencias existentes entre el mejor algoritmo encontrado (denominado como algoritmo de

control) y el resto. Estos test se limitan a detectar la existencia o no de diferencias en todo el conjunto de resultados. Si las encuentran, es necesario proceder con un procedimiento post-hoc de identificación de diferencias (basados en la distribución normal).

Para aproximar el estimador z de una distribución normal a partir de la diferencias entre dos rankings para cada uno de los tests descritos anteriormente se debe realizar lo siguiente:

- Test de Friedman: En [4] podemos ver que la expresión para calcular z es

$$z = (R_i - R_j)/\sqrt{\frac{k(k+1)}{6n}} \quad (11)$$

donde R_i y R_j son los rankings medios obtenidos por el test de Friedman.

- Test de Friedman Alineado: Puesto que los rankings relativos se convierten en absolutos, la expresión para calcular el valor z es la misma que la que se usa en el test de Kruskal-Wallis [4]

$$z = (\hat{R}_i - \hat{R}_j)/\sqrt{\frac{k(k+1)}{6n}} \quad (12)$$

donde $\hat{R}_i$ y $\hat{R}_j$ son los rankings medios obtenidos por el test de Friedman.

- Test de Quade: En [3], el estadístico para comparar dos algoritmos se proporciona usando la distribución t de Student, pero es posible aproximarla a la distribución normal para calcular el valor z [14].

$$z = (T_i - T_j)/\sqrt{\frac{k(k+1)(2n+1)(k-1)}{18n(n+1)}} \quad (13)$$

donde $T_i = \frac{W_i}{n(n+1)/2}$, $T_j = \frac{W_j}{n(n+1)/2}$ y W_i y W_i a son los rankings sin ajuste de medias proporcionados por el Test de Quade.

A continuación, presentamos los 4 procedimientos más útiles para detectar diferencias a posteriori. Existen más alternativas, pero apenas ofrecen alguna diferencia de potencial con respecto a las 4 que se detallan aquí [10]. Una vez obtenido el valor z, es posible obtener a partir de él el p-valor no ajustado correspondiente. La utilidad de estos procedimientos *post-hoc* radica en que son capaces de calcular el P-Valor Ajustado (PVA) considerando la familia de hipótesis completa para cada pareja de algoritmos comparados.

Seguidamente, se detalla la manera de calcular el PVA para cada procedimiento [16]. Para ello, usaremos la siguiente notación:

- Los índices i y j corresponden a una hipótesis de comparación entre los algoritmos i y j, de acuerdo a un orden incremental de sus p-valores no ajustados p. El índice i siempre se refiere al algoritmo de control cuyo PVA será calculado.
- p_j es el valor p obtenido para la hipótesis j-ésima.
- k es el número de algoritmos a comparar.

TABLA IV
P-VALORES AJUSTADOS USANDO **DE-EXP** COMO MÉTODO DE CONTROL.

i	1	2	3
Algoritmo	**PSO**	**SSGA**	**SS-BLX**
	Friedman		
p no ajustado	0.000005	0.021424	0.089519
Bonferroni PVA	0.000016	0.064271	0.268557
Holm PVA	0.000016	0.042847	0.089519
Hochberg PVA	0.000016	0.042847	0.089519
Li PVA	0.000006	0.022989	0.089519
	Friedman Alineado		
p no ajustado	0.000461	0.022818	0.233318
Bonferroni PVA	0.001383	0.068454	0.699953
Holm PVA	0.001383	0.045636	0.233318
Hochberg PVA	0.001383	0.045636	0.233318
Li PVA	0.000601	0.028902	0.233318
	Quade		
p no ajustado	0.002025	0.033683	0.135671
Bonferroni PVA	0.006076	0.101048	0.407012
Holm PVA	0.006076	0.067366	0.135671
Hochberg PVA	0.006076	0.067366	0.135671
Li PVA	0.002338	0.037508	0.135671

Los 4 procedimientos pueden describirse según su modo de calcular los PVAs como sigue:

- Bonferroni PVA_i: $\min\{v; 1\}$ donde $v = (k-1) \cdot p_i$.
- Holm PVA_i: $\min\{v; 1\}$ donde $v = \max\{(k-j) \cdot p_j : 1 \leq j \leq i\}$.
- Hochberg PVA_i: $\max\{(k-j) \cdot p_j : (k-1) \geq j \geq i\}$.
- Li PVA_i: $p_i/(p_i + 1 - p_{k-1})$.

La Tabla IV muestra los PVAs para el marco experimental considerado en este trabajo. Como puede verse, este ejemplo es válido para observar ligeras diferencias de potencia entre los procedimientos estudiados. Si comparamos los PVAs obtenidos con el nivel de significancia que se establece a priori en todo estudio estadístico, podremos contabilizar el número de rechazos (si $\text{VAP}_i \leq \alpha$) y retenciones de hipótesis que se obtienen.

V. Otros tests para comparaciones multiples

Además de los tests de Friedman, Friedman Alineado y Quade, existen otros procedimientos capaces de evaluar el rendimiento de varios algoritmos de forma simultánea, aportando información de interés sobre la comparación. En esta sección se revisarán dos de estos métodos: El Test de Signos Múltiple y la Estimación de Contraste.

A. Test de Signos Múltiple

El Test de Signos Múltiple [15] es un sencillo procedimiento para realizar comparaciones $1xN$. Aunque es menos potente que los tests anteriores, puede ser útil en casos en los que se desee realizar un primer estudio preliminar de los resultados.

Este test es una extensión del Test de Signos convencional, capaz de comparar simultáneamente k métodos contra un método de control. Consiste en los siguientes pasos:
1. Representar con x_{i1} los valores de rendimiento del método control en el conjunto i-ésimo. Representar con x_{ij} los valores de rendimiento del resto de métodos (j-ésimo método en el conjunto i-ésimo).
2. Calcular los signos de las diferencias $d_{ij} = x_{i1} - x_{ij}$, asignando (+) en el caso de que el método de control ofrezca un peor rendimiento, o (-) en caso contrario. Las diferencias a 0 (empates, (=)) se descartan.
3. Calcular r_j como el número de diferencias d_{ij} con el signo menos frecuente, (+) ó (-), dentro de un emparejamiento del algoritmo j con el control.
4. Sea M_1 la respuesta mediana de una muestra de resultados del algoritmo de control y M_j la respuestas mediana de una muestra de resultados del algoritmo j-ésimo. El test de signos permite aplicar una de las dos reglas de decisión siguientes:

- Para comprobar $H0 : Mj \geq M1$ contra $H0 : Mj < M1$, se rechaza H0 si el número de signos (+) es igual o inferior que el valor crítico de R_j que aparece en la Tabla A.1 de [10] para $k-1$ (número de algoritmos excluyendo el control), n (número de problemas) y el nivel de significancia escogido.
- Para comprobar $H0 : Mj \leq M1$ contra $H0 : Mj > M1$, se rechaza H0 si el número de signos (-) es igual o inferior que el valor crítico de R_j que aparece en la Tabla A.1 de [10] para $k-1$ (número de algoritmos excluyendo el control), n (número de problemas) y el nivel de significancia escogido.

La Tabla V muestra las cálculos realizados por este procedimiento. Usaremos dos niveles de significancia $\alpha = 0{,}1$ y $\alpha = 0{,}05$. Las hipótesis son $H0 : Mj \geq M1$ contra $H0 : Mj < M1$; esto es, el algoritmo de control **DE-EXP** es mejor que el resto de algoritmos. Tenemos que $k-1 = 3$, $n = 23$ en el caso del algoritmo **SSGA** (existen 2 empates (=) que se descartan) y $n = 24$ en el resto de casos (se descarta 1 empate (=) en cada uno). La Tabla A.1 de [10] indica que con $\alpha = 0{,}1$ los valores críticos R_j son 7 y 6 para $n = 24$ y $n = 23$, respectivamente. Con $\alpha = 0{,}1$, el valor crítico es 6 en todos los casos.

Puesto que el número de (+) de **PSO** y **SSGA** es igual o inferior a ambos valores críticos, puede establecerse que **DE-EXP** presenta un mejor rendimiento que ellos (con $\alpha = 0{,}05$). Sin embargo, no puede decirse lo mismo de la comparación con **SS-BLX**, por lo que se concluye que muestran un rendimiento equivalente.

TABLA V
TEST DE SIGNOS MÚLTIPLE USANDO **DE-EXP** COMO MÉTODO DE CONTROL.

Algoritmo	**DE-EXP**	**PSO**	**SSGA**	**SS-BLX**
Orden	1 (control)	2	3	4
F1	8,26E-06	1,23E-01 (-)	8,42E-06 (-)	3,40E+02 (-)
F2	8,18E-06	2,60E+01 (-)	8,72E-02 (-)	1,73E+03 (-)
F3	9,94E+04	5,17E+07 (-)	7,95E+07 (-)	1,84E+08 (-)
F4	8,35E-06	2,49E+03 (-)	2,59E+00 (-)	6,23E+03 (-)
F5	8,51E-06	4,10E+05 (-)	1,34E+05 (-)	2,19E+03 (-)
F6	8,39E-06	7,31E+05 (-)	6,17E+03 (-)	1,15E+05 (-)
F7	1,27E+06	2,68E+02 (+)	1,27E+06 (=)	1,97E+06 (-)
F8	2,04E+04	2,04E+04 (=)	2,04E+04 (=)	2,04E+04 (=)
F9	8,15E-06	1,44E+04 (-)	7,29E-06 (+)	4,20E+03 (-)
F10	1,12E+04	1,40E+04 (-)	1,71E+04 (-)	1,24E+04 (-)
F11	2,07E+03	5,59E+03 (-)	3,26E+03 (-)	2,93E+03 (-)
F12	6,31E+04	6,36E+05 (-)	2,79E+05 (-)	1,51E+05 (-)
F13	6,40E+02	1,50E+03 (-)	6,71E+02 (-)	3,25E+02 (+)
F14	3,16E+03	3,30E+03 (-)	2,26E+03 (+)	2,80E+03 (+)
F15	2,94E+05	3,40E+05 (-)	2,92E+05 (+)	1,14E+05 (+)
F16	1,13E+05	1,33E+05 (-)	1,05E+05 (+)	1,04E+05 (+)
F17	1,31E+05	1,50E+05 (-)	1,19E+05 (+)	1,18E+05 (+)
F18	4,48E+05	8,51E+05 (-)	8,06E+05 (-)	7,67E+05 (-)
F19	4,34E+05	8,50E+05 (-)	8,90E+05 (-)	7,56E+05 (-)
F20	4,19E+05	8,51E+05 (-)	8,89E+05 (-)	7,46E+05 (-)
F21	5,42E+05	9,14E+05 (-)	8,52E+05 (-)	4,85E+05 (+)
F22	7,72E+05	8,07E+05 (-)	7,52E+05 (+)	6,83E+05 (+)
F23	5,82E+05	1,03E+06 (-)	1,00E+06 (-)	5,74E+05 (+)
F24	2,02E+05	4,12E+05 (-)	2,36E+05 (-)	2,51E+05 (-)
F25	1,74E+06	5,10E+05 (+)	1,75E+06 (-)	1,79E+06 (-)
# (+)	-	2	6	8
# (-)	-	22	17	16
$R_j\alpha = 0{,}1$	-	7	6	7
$R_j\alpha = 0{,}05$	-	6	6	6

B. Estimación de Contraste

Utilizando los datos resultantes de la ejecución de varios algoritmos sobre múltiples problemas, un investigador podría estar interesado en la estimación de las diferencias entre el rendimiento de dos de ellos. Un procedimiento que busca este propósito es la Estimación de Contraste [6], la cual supone que las diferencias esperadas entre rendimientos de algoritmos son las mismas entre problemas. Asumimos que cada medición de rendimiento se refleja como diferencias entre rendimientos de los algoritmos. Es decir, estamos interesados en estimar el contraste entre medianas de muestras de resultados considerando todas las comparaciones por pares. Para ello, la Estimación de Contraste obtiene una diferencia cuantitativa calculada mediante medianas entre dos algoritmos.

Se procede de la siguiente manera:

1. Para cada pareja de k algoritmos en el experimento, calculamos la diferencia entre los rendimientos de ambos en cada uno de los n problemas, $D_i(uv) = x_{iu} - x_{iv}$ donde $i = 1 \ldots n$, $u = 1 \ldots k$ y $v = 1 \ldots k$. Se consideran solo aquellas diferencias donde $u < v$.
2. Buscamos la mediana de cada conjunto de diferencias, Z_{uv}. Z_{uv} es el *estimador no ajustado* de la diferencia entre medianas de u y v. Nótese que $Z_{uu} = 0$.
3. Calculamos la media de cada conjunto de medianas no ajustadas que tienen el mismo prefijo u , m_u:

$$m_u = \frac{\sum_{j=1}^{k} Z_{uv}}{k}, u = 1, \ldots, k \qquad (14)$$

4. El estimador de la diferencia entre medianas de u y v es $m_u - m_v$, donde u y v son un par cualquiera de los algoritmos a comparar.

Este procedimiento se ilustra considerando el mismo marco experimental que en los estudios previos. La Tabla VI muestra los cálculos realizados.

TABLA VI
DIFERENCIAS ENTRE PARES DE RENDIMIENTOS EN CADA PROBLEMA PARA CADA PAR DE ALGORITMOS.

Dif.	$D_{i(12)}$	$D_{i(13)}$	$D_{i(14)}$	$D_{i(23)}$	$D_{i(24)}$	$D_{i(34)}$
F1	1,23E-01	-3,40E+02	1,23E-01	-3,40E+02	1,60E-07	3,40E+02
F2	2,59E+01	-1,70E+03	2,59E+01	-1,73E+03	8,72E-02	1,73E+03
F3	-2,77E+07	-1,33E+08	5,16E+07	-1,05E+08	7,94E+07	1,84E+08
F4	2,49E+03	-3,74E+03	2,49E+03	-6,23E+03	2,58E+00	6,23E+03
F5	2,75E+05	4,07E+05	4,09E+05	1,32E+05	1,34E+05	2,18E+03
F6	7,25E+05	6,17E+05	7,31E+05	-1,08E+05	6,17E+03	1,14E+05
F7	-1,27E+06	-1,97E+06	-1,26E+06	-6,95E+05	6,00E+03	7,01E+05
F8	6,00E+01	8,00E+01	5,00E+01	2,00E+01	-1,00E+01	-3,00E+01
F9	1,44E+04	1,02E+04	1,44E+04	-4,19E+03	-8,65E-07	4,19E+03
F10	-3,08E+03	1,65E+03	2,86E+03	4,73E+03	5,94E+03	1,21E+03
F11	2,34E+03	2,66E+03	3,52E+03	3,26E+02	1,19E+03	8,62E+02
F12	3,57E+05	4,86E+05	5,73E+05	1,29E+05	2,16E+05	8,75E+04
F13	8,32E+02	1,18E+03	8,63E+02	3,47E+02	3,10E+01	-3,16E+02
F14	1,04E+03	5,08E+02	1,46E+02	-5,32E+02	-8,94E+02	-3,62E+02
F15	4,78E+04	2,26E+05	4,58E+04	1,78E+05	-2,00E+03	-1,80E+05
F16	2,80E+04	2,92E+04	2,08E+04	1,20E+03	-7,20E+03	-8,40E+03
F17	3,12E+04	3,14E+04	1,85E+04	2,00E+02	-1,27E+04	-1,29E+04
F18	4,49E+04	8,44E+04	4,03E+05	3,95E+04	3,58E+05	3,19E+05
F19	-4,02E+04	9,42E+04	4,16E+05	1,34E+05	4,56E+05	3,21E+05
F20	-3,84E+04	1,05E+05	4,32E+05	1,43E+05	4,71E+05	3,28E+05
F21	6,16E+04	4,29E+05	3,72E+05	3,67E+05	3,10E+05	-5,69E+04
F22	5,52E+04	1,24E+05	3,51E+04	6,91E+04	-2,01E+04	-8,92E+04
F23	2,40E+04	4,54E+05	4,46E+05	4,30E+05	4,22E+05	-8,40E+03
F24	1,76E+05	1,61E+05	2,10E+05	-1,53E+04	3,40E+04	4,93E+04
F25	-1,24E+06	-1,28E+06	-1,23E+06	-4,70E+04	5,00E+03	5,20E+04
Med.	2,49E+03	2,92E+04	2,08E+04	3,47E+02	1,19E+03	1,73E+03

TABLA VII
ESTIMACIÓN DE CONTRASTE BASADA EN MEDIANAS ENTRE TODOS LOS ALGORITMOS DEL ESTUDIO EXPERIMENTAL.

	PSO	SSGA	SS BLX	DE-EXP
PSO	0	1,31E+04	1,98E+04	1,86E+04
SSGA	-1,31E+04	0	6,67E+03	5,49E+03
SS BLX	-1,98E+04	-6,67E+03	0	-1,17E+03
DE-EXP	-1,86E+04	-5,49E+03	1,17E+03	0

Una vez obtenidas las medianas Z_{uv} con la Tabla VI calculamos las medias para m_1 y m_2:

$$m_1 = \frac{2,49E+03 + 2,92E+04 + 2,08E+04}{4} = 1,31E+04$$

$$m_2 = \frac{-2,49E+03 + 3,47E+02 + 1,19E+03}{4} = -9,53E+02$$

Nuestro estimador de la diferencia entre medianas $m_1 - m_2$ es $1,31E+04-(-9,53E+02) = 1,41E+04$. Es decir, la diferencia en error entre PSO y SSGA estimada sobre multiples problemas es igual a $1,41E+04$ (el resultado correcto varía ligeramente debido al error cometido en la Tabla VI al tomar solo 3 dígitos significativos. El valor exacto de esta estimación es de $1,31E+04$).

La Tabla VII muestra los estimadores obtenidos entre todos los pares de algoritmos.

VI. PREGUNTAS FRECUENTES

En esta sección se recogen y responden una serie de preguntas frecuentes acerca del uso de tests no paramétricos.

- **¿Podemos analizar cualquier medida de rendimiento?** Con estadística no paramétrica se puede analizar cualquier medida unaria (asociadas a un solo algoritmo) que tenga un rango de salida definido. Este rango no tiene por qué estar limitado, por lo que es factible analizar tiempos o incluso requisitos de memoria.
- **¿Podemos comparar algoritmos determinísticos con estocásticos?** Los tests no paramétricos pueden comparar ambos tipos de algoritmos simultáneamente porque pueden ser aplicados en comparaciones multi-dominio, donde la muestra de resultados la componen un resultado de rendimiento por algoritmo y dominio.
- **¿Cómo se han de obtener los resultados de cada algoritmo?** Esta cuestión no concierne al uso de estadística no paramétrica, puesto que estos tests estadísticos necesitan un resultado por pareja algoritmo-dominio. La obtención de dicho resultado debe seguir un procedimiento conocido y estandarizado por todos los algoritmos, como puede ser el empleo de técnicas de validación y utilizar resultados medios o medianos provenientes de un número suficientemente elevado de ejecuciones en algoritmos probabilísticos.
- **¿Qué relación debe haber entre el número de algoritmos y el número de conjuntos de datos para hacer un análisis estadístico adecuado?** En la comparación de nuevas propuestas con varios algoritmos, el número de problemas (dominios) debe ser superior al doble de algoritmos como mínimo. Con menos problemas, es altamente probable que no se pueda rechazar ninguna hipótesis.
- **¿Existe un tope de conjuntos de datos que podamos utilizar?** No existe un tope teórico, aunque si el número de problemas es muy grande en relación al de algoritmos, los resultados tienden a ser ineficaces según el teorema central de límite [14]. Para comparaciones múltiples con un método control, dependerá del número de algoritmos que se comparen, aunque podríamos indicar que utilizar más de $n > 8 \cdot k$ conjuntos de datos podría ser excesivo y resultar en comparaciones no significativas.
- **El Test de Wilcoxon aplicado varias veces funciona mejor que un test de múltiples comparaciones ¿Es válido en estos casos?** El Test de Wilcoxon se puede utilizar siguiendo un enfoque de múltiples comparaciones pero los resultados obtenidos no se pueden considerar en familia. En el momento en que se hace más de una comparación con el Test de Wilcoxon, el nivel de significancia establecido a priori puede superarse porque no se controla el error producido en una familia. Es precisamente en este aspecto donde cobran especial relevancia los tests de comparaciones múltiples.
- **¿Podemos usar solo los valores de rankings obtenidos para justificar los resultados?** Con los valores de rankings obtenidos con el Test de Friedman y derivados podemos establecer una ordenación clara entre algoritmos e incluso medir las diferencias entre ellos, pero no se puede concluir que una propuesta es mejor que otra a no ser que la hipótesis de comparación quede rechazada.
- **¿Es necesario comprobar que la hipótesis nula es rechazada por el Test de Friedman y derivados antes de proceder al análisis de**

comparaciones posterior? Es conveniente, aunque por definición, se pueden calcular de forma independiente.

▪ **¿Cuando es recomendable usar el Test de Friedman Alineado o el Test de Quade en vez del clásico Test de Friedman?** Las diferencias entre los tres métodos son pequeñas y muy dependientes de los resultados a analizar. Ciertos estudios teóricos demuestran que tanto el Test de Friedman Alineado como el Test de Quade tienen mejor rendimiento cuando comparamos pocos algoritmos (no más de 4). El Test de Quade también supone un cierto riesgo porque asume que los problemas más relevantes son aquellos que presentan mayores diferencias entre los métodos, y esto no tiene por qué ser así siempre.

▪ **¿Qué procedimientos post-hoc deberían usarse?** Consideramos que el Test de Holm debe aparecer siempre en toda comparación, mientras que el Test de Bonferroni nunca por su conservatividad. El Test de Hochberg y el Test de Li pueden servir de complemento cuando su uso permita rechazar más hipótesis de las que rechaza el Test de Holm. Toda hipótesis rechazada por cualquier procedimiento está correctamente rechazada, puesto que todos los tests a posteriori ofrecen un control fuerte de la tasa de error en familia. Sin embargo hay tests, como el de Li, que están influenciados por los p-valores no ajustados obtenidos en las hipótesis iniciales, y sólo cuando son menores que 0,5 el test obtiene su mejor rendimiento.

▪ **¿Qué utilidad tiene el Test de Signos Multiple?** El Test de Signos Multiple puede considerarse como una alternativa simple a la hora de realizar un test de comparaciones múltiples con un algoritmo de control. Tal y como hemos mostrado, se trata de un procedimiento rápido y sencillo de aplicar, pero con una capacidad de rechazo inferior a la del Test de Friedman. Generalmente, se recomienda su uso cuando las diferencias encontradas entre el método de control y el resto de algoritmos con respecto a alguna medida de rendimiento sean muy claras [10].

▪ **¿En qué casos puede ser interesante emplear la Estimación de Contraste?** La Estimación de Contraste basada en medianas proporciona una forma de calcular las diferencias de rendimiento entre dos algoritmos, dando especial importancia a la mediana de sus resultados. Dado la dificultad de empleo de los Test Paramétricos para contrastar experimentos [9], la Estimación de Contraste aparece como un métrica eficaz y fiable para evaluar las diferencias entre varios algoritmos en entornos multiproblema.

VII. Conclusiones

En este trabajo hemos presentado una familia de tests no paramétricos para comparar nuevas algoritmos, incluyendo ejemplos y recomendaciones sobre su uso. En el sitio web temático de SCI2S sobre Inferencia Estadística en Inteligencia Computacional y Minería de Datos (`http://sci2s.ugr.es/sicidm`) se proporciona información más avanzada sobre el tema y varios paquetes de software para realizar los análisis estadísticos descritos en este trabajo. Asimismo, el proyecto KEEL (`http://www.keel.es`) [1], [2] ofrece una completa herramienta de análisis estadístico no paramétrico, capaz de realizar todos los análisis destacados en este trabajo.

Agradecimientos

Subvencionado por los proyectos TIN2011-28488 y TIC-2010-6858. J. Derrac disfruta de una beca FPU del Ministerio de Educación y Ciencia.

Referencias

[1] J. Alcalá-Fdez, L. Sánchez, S. García, M.J. del Jesus, S. Ventura, J.M. Garrell, J. Otero, C. Romero, J. Bacardit, V.M. Rivas, J.C. Fernández, F. Herrera, *KEEL: a software tool to assess evolutionary algorithms for data mining problems*, Soft Computing, vol. 13, no. 3, pp. 307-318, 2008.

[2] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, F. Herrera, *KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework*, Journal of Multiple-Valued Logic and Soft Computing, vol. 17, no. 2-3, pp. 255-287, 2011.

[3] W. J. Conover, *Practical Nonparametric Statistics*, John Wiley and Sons, 1999.

[4] W. W. Daniel, *Applied Nonparametric Statistics*, Duxbury Thomson Learning, 1990.

[5] J. Derrac, S. García, D. Molina, F. Herrera, *A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms*, Swarm and Evolutionary Computation, vol. 1, no. 1, pp. 3-18, 2011.

[6] K. Doksum, *Robust procedures for some linear models with one observation per cell*, Annals of Mathematical Statistics, vol. 38, pp. 878-883, 1967.

[7] M. Friedman, *The use of ranks to avoid the assumption of normality implicit in the analysis of variance*, Journal of the American Statistical Association, vol. 32, pp. 674-701, 1937.

[8] M. Friedman, *A comparison of alternative tests of significance for the problem of m rankings*, Annals of Mathematical Statistics, vol. 11, pp. 86-92, 1940.

[9] S. García, D. Molina, M. Lozano, F. Herrera, *A study on the use of non-parametric tests for analyzing the evolutionary algorithms' behaviour: A case study on the CEC'2005 special session on real parameter optimization*, Journal of Heuristics, vol. 15, pp. 617-644, 2009.

[10] S. García, A. Fernández, J. Luengo, F. Herrera, *Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power*, Information Sciences, vol. 180, pp. 2044-2064, 2010.

[11] J. Hodges, E. Lehmann, *Ranks methods for combination of independent experiments in analysis of variance*, Annals of Mathematical Statistics, vol. 33, pp. 482-497, 1962.

[12] J.D. Gibbons, S. Chakraborti, *Nonparametric Statistical Inference, 5th Edition*, Chapman & Hall, 2010.

[13] D. Quade, *Using weighted rankings in the analysis of complete blocks with additive block effects*, Journal of the American Statistical Association, vol. 74, pp. 680-683, 1979.

[14] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures, 5th Edition*, Chapman & Hall/CRC, 2011.

[15] R. Steel, *A multiple comparison sign test: treatments versus control*, Journal of the American Statistical Association, vol. 54, pp. 767-775, 1959.

[16] S.Y.P.H. Westfall, *Resampling-based Multiple Testing: Examples and Methods for p-Value Adjustment*, John Wiley and Sons, 2004.