

# Sobre la comparación de metaheurísticas mediante técnicas estadísticas no paramétricas

José A. Moreno Pérez<sup>(1)</sup> Clara Campos Rodríguez<sup>(1)</sup> Manuel Laguna<sup>(2)</sup>

*Resumen*— En este trabajo se describen los fundamentos para la comparación de heurísticas mediante técnicas estadísticas incluyendo los contrastes disponibles y las características que los hacen apropiados para las circunstancias más usuales. Se propone el uso de unos contrastes no paramétricos sencillos, seguros y robustos para la comparación estadística de metaheurísticas sobre un conjunto de instancias: el test de los rangos con signo de Wilcoxon para la comparación de dos heurísticas y, para la comparación de más de dos heurísticas, el test de Friedman de comparaciones múltiples con los correspondientes test a posteriori de Namenyi y de Bonferroni-Dunn para establecer las conclusiones mediante los procedimientos de Holm y de Hochberg. Se ejemplifica el uso de estos contrastes sobre unos datos supuestos.

*Palabras clave*— Comparación de Heurísticas, Contrastes no paramétricos.

## I. INTRODUCCIÓN

En los últimos años, la comunidad de Metaheurística ha alcanzado un claro convencimiento de la necesidad de aplicar técnicas estadísticas para validar los avances conseguidos. Esto es reflejo de cierta madurez del área, del crecimiento continuo de la capacidad de cómputo, del incremento de las aplicaciones reales y de la disponibilidad de cada vez más metaheurísticas. Las características del campo facilitan el desarrollo e implementación de nuevas metaheurísticas o la modificación de las existentes, y la realización de experimentos para establecer comparaciones entre ellas.

En un artículo típico sobre metaheurísticas, se propone un nuevo algoritmo metaheurístico (frecuentemente un híbrido) o una versión de una metaheurística ya conocida; o se introduce una componente, o un nuevo paso (de pre-procesamiento, de post-optimización o intermedio); y se realiza, explícita o implícitamente, la hipótesis de que tal desarrollo mejora el rendimiento de los algoritmos previamente existentes. En otros artículos, se proponen diversas soluciones heurísticas alternativas para un problema y el propósito es analizar cuál resuelve el problema con éxito y cuál fracasa. Para la parte experimental del estudio se selecciona un conjunto de instancias, reales o generadas aleatoriamente, y sobre ellas se ejecutan los algoritmos y se mide el rendimiento. El trabajo suele incluir un conjunto de tablas con los indicadores del rendimiento alcanzados por los algoritmos participantes en el estudio y, con algún proced-

imiento de sentido común sobre la visualización de esos indicadores o con la aplicación de alguna técnica estadística, se determina si las diferencias observadas pueden atribuirse al azar o son evidencia suficiente de una diferencia real en el rendimiento de los algoritmos. A partir de este análisis se realizan conclusiones en el sentido del objetivo propuesto, que se suele concretar en que la propuesta realizada mejora significativamente las anteriores. En otros estudios se justifica la contribución del método propuesto con una clara mejora en otras propiedades menos cuantificables (simplicidad en el algoritmo o su implementación, menor número de parámetros a ajustar, inspiración en la naturaleza) sin un empeoramiento significativo en el indicador del rendimiento. Todas estas conclusiones deben estar sustentadas en contrastes estadísticos rigurosos aplicados con imparcialidad en lugar de en la mera observación de tablas numéricas de cierto tamaño.

Este tipo de situación, con sus propias particularidades, se viene presentando desde hace muchísimo tiempo en las investigaciones de laboratorio de las ciencias experimentales clásicas y en las pruebas de los diseños prácticos realizados desde la ingeniería. Por esto empezó a desarrollarse a principios del siglo XX la teoría de los contrastes de hipótesis estadísticas para realizar un planteamiento objetivo y fiable de las investigaciones. Se ha estudiado la forma de diseñar los test apropiados para realizar los contrastes en las circunstancias más usuales para conseguir el máximo rendimiento de los datos disponibles en la obtención de conclusiones con las correspondientes garantías.

En las siguientes secciones se analizan algunas cuestiones de la elección del indicador del rendimiento y del planteamiento general de los contrastes de hipótesis estadísticas para comparar heurísticas. Se describen los contrastes de hipótesis paramétricos y no paramétricos para comparar dos o más metaheurísticas sobre un conjunto de instancias. Se analizan las principales características que los diferencian: que suposiciones están detrás de cada test, qué es lo que contrastan realmente y qué circunstancias los hacen especialmente apropiados. El trabajo finaliza con unas breves conclusiones.

## II. RENDIMIENTO DE LAS METAHEURÍSTICAS

La comparación entre las metaheurísticas debe hacerse en términos del mayor o menor cumplimiento de las propiedades deseables; aquellas que favorecen el interés práctico y teórico de las meta-

<sup>(1)</sup> Instituto Universitario de Desarrollo Regional. Universidad de La Laguna. 38271 La Laguna. e-mail: jamoreno@ull.es.

<sup>(2)</sup> Universidad de Colorado at Boulder, Colorado. USA. e-mail: laguna@colorado.edu.

heurísticas (ver [14]). Indican direcciones a las que dirigir los esfuerzos para contribuir al desarrollo del área, pero no será posible mejorar todas las propiedades a la vez. Unas propiedades son parcialmente contrapuestas pero otras están muy relacionadas entre sí y apuntan conjuntamente en un mismo sentido. Esto permite agruparlas en torno a ciertas características, como la facilidad de su *comprensión* y aplicación (simplicidad, precisión y coherencia), las que propician su amplia *aplicabilidad* (generalidad, adaptabilidad y robustez), las que favorecen su utilidad en Sistemas de Ayuda a la Decisión (interactividad, multiplicidad y autonomía) y, finalmente, las que intervendrían en la evaluación de su *rendimiento* práctico (eficiencia, efectividad y eficacia).

Las propiedades cuantificables que pueden intervenir en la evaluación del rendimiento son la eficiencia, la eficacia, la efectividad, y tal vez la robustez. Estas propiedades están muy relacionadas y aunque frecuentemente no se establece una separación clara conviene distinguirlas y establecer precisamente qué miden. La *eficiencia* hace referencia a la cantidad de recursos empleados (espacio y, principalmente, tiempo) al actuar, la *eficacia* se relaciona con la probabilidad de alcanzar una solución óptima y la *efectividad* con la calidad de las soluciones propuestas. La *robustez* refleja la variabilidad de comportamiento al modificarse las características de las instancias sobre las que se ejecuta.

Las magnitudes asociadas a estas propiedades pueden ser fácilmente evaluadas en una ejecución de la metaheurística y permiten establecer una fórmula de computar un índice del rendimiento sobre una serie de casos para, en base a ellos, establecer comparaciones. Las cantidades usadas pueden ser el valor de la función objetivo alcanzada o, si se conoce el valor óptimo, el tiempo empleado en alcanzarlo, o si no se alcanza siempre, el número de veces que se alcanza en un tiempo razonable, o la razón con respecto al óptimo de la solución aportada. En el indicador del rendimiento adoptado se puede hacer intervenir una de ellas, o una combinación razonable de varias.

Para aplicar los contrastes estadísticos usuales en las ciencias experimentales sólo se requiere que los indicadores sean apropiados para comparar el rendimiento de los algoritmos sobre cada instancia, en el sentido de que un mayor valor de dicho indicador es señal de un mejor rendimiento del algoritmo. Las conclusiones que estarían soportadas por las técnicas estadísticas son las de una mejora significativa de alguno de estos indicadores o las de inexistencia de un empeoramiento significativo. Estos y otros aspectos de las cantidades medidas en los experimentos han sido analizadas en la literatura ([2]).

### III. CONTRASTES DE HIPÓTESIS

La teoría y la práctica del contraste de hipótesis estadístico surge a finales del siglo XIX gracias fun-

damentalmente a los trabajos de R.A. Fisher con el siguiente planteamiento general [6]. En la situación de partida de un estudio experimental se plantea una *hipótesis nula*, denotada  $H_0$ , que, como su nombre indica, debe representar la situación previamente establecida o aquella que implica que las propuestas contempladas en el estudio no suponen novedad o mejora significativa, frente a la hipótesis contraria o *hipótesis alternativa*, denotada  $H_1$ . Esta hipótesis alternativa puede ser tan general como el no cumplimiento completo de toda la hipótesis nula, pero, apoyada o no en argumentos teóricos en la dirección deseada, viene a representar de forma muy concreta, los planteamientos o conclusiones que se persigue sustentar en el estudio estadístico. En nuestro contexto del estudio experimental del rendimiento de metaheurísticas, la hipótesis nula corresponde con la idea de que la nueva heurística que se propone no supone mejora con respecto a la solución estándar o las ya conocidas con las que se trata de comparar, o, en el caso de que se esté realizando un estudio para comparar entre sí varias propuestas heurísticas, corresponde con que las diferencias observadas en su comportamiento se pueden atribuir al efecto del azar. En el primero de estos casos, la hipótesis alternativa sería que la nueva propuesta heurística realizada supone una mejora, pudiéndose llegar a concretar una medida de dicha mejora, y en el segundo caso, que existe alguna diferencia, pudiéndose también concretar dónde (entre qué par de propuestas) radica exactamente la diferencia e incluso alguna medida de dicha diferencia.

Para contrastar estadísticamente una hipótesis nula  $H_0$ , frente a la hipótesis alternativa  $H_1$ , se construye un *test de hipótesis* consistente en unas formulas o mecanismo para, a partir de los datos obtenidos, optar por rechazar o no rechazar la hipótesis nula, aceptando la alternativa. En la práctica los tests de hipótesis se diseñan utilizando una o más variables aleatorias cuyo comportamiento dependa de que sea cierta la hipótesis nula o la alternativa. Estas variables se obtienen de alguna fórmula aplicada a los datos obtenidos directamente de la experimentación, como son la media, la desviación típica o fórmulas similares, que se denominan *estadísticos*. Estos estadísticos deben resumir los datos pero construidos de forma que se extraiga la información contenida en los datos que afecte a las hipótesis nula y alternativa, para lo que existe una parte de la estadística que se ocupa de ello. Una vez seleccionado el estadístico  $T$  en el que basar el test se determina un conjunto  $R$  de posibles valores del estadístico, llamado *región crítica*, para que, en el caso de que el valor del estadístico  $T$  esté en la región crítica se opte por rechazar la hipótesis nula en favor de la hipótesis alternativa, que es aceptada. En caso contrario, no se rechaza la hipótesis nula y por tanto no se acepta la alternativa. Nótese que por la diferente naturaleza de ambas hipótesis con respecto a los objetivos de la

investigación, se habla de rechazar o no la hipótesis nula y de aceptar o no la hipótesis alternativa, evitando otras expresiones. Conociendo cual debe ser el comportamiento del estadístico  $T$  dependiendo de si es cierta la hipótesis nula o la alternativa, la región crítica se construye con aquellos valores que son muy verosímiles para el estadístico si es cierta la hipótesis alternativa y muy poco verosímiles si es cierta la hipótesis nula. Teniendo en cuenta el planteamiento original del contraste de hipótesis se pueden cometer básicamente dos errores: el denominado *error de tipo I* consistente en rechazar la hipótesis nula siendo cierta y el *error de tipo II* consistente en no rechazarla siendo cierta la hipótesis alternativa. Con un test de hipótesis basado en el estadístico  $T$  y la región crítica  $R$ , la probabilidad de cometer el error de tipo I es  $Pr(T \in R|H_0)$  y la probabilidad de cometer el error de tipo II es  $Pr(T \notin R|H_1)$ .

El *nivel de confianza* del test establece una cota para la probabilidad,  $Pr(T \in R|H_0)$ , de cometer el error de tipo I y la probabilidad,  $Pr(T \in R|H_1)$ , de no cometer el error de tipo II se denomina poder o *potencia* del test. El nivel de confianza se fija mediante unos porcentajes próximos al 100% y se denota por  $1 - \alpha$ . Los valores usuales para  $\alpha$ , denominado *nivel de significación*, suelen ser de 0,01, 0,05 o 0,10 que expresados en porcentajes son el 1%, el 5% y el 10% y corresponden, respectivamente a niveles de confianza del 99%, del 95% y del 90%. El propósito ideal con el diseño del test mediante la elección de  $T$  y  $R$  es minimizar las dos probabilidades de error; sin embargo, dado que esto no es posible en general, se fija el nivel de significación  $\alpha$  en un valor apropiado y se trata de que la potencia, denotada por  $\beta$ , sea máxima.

En muchos casos, las particularidades del problema permiten garantizar que la región crítica apropiada consiste en los valores del estadístico que sobrepasan un valor determinado, el valor crítico. Por tanto, la construcción del test consiste en, dado el valor de  $\alpha$ , obtener el valor crítico  $t_\alpha$ , tal que  $Pr(T > t_\alpha|H_0) = \alpha$ . Una vez realizado el experimento, evaluando el estadístico  $T$  en los datos se obtiene el valor práctico  $t$ , entonces si  $t > t_\alpha$  se rechaza la hipótesis nula porque  $Pr(T > t|H_0) \leq \alpha$ . Sin embargo, una información más precisa del nivel de confianza con el que se rechaza la hipótesis nula, nos la da el cálculo preciso de esta probabilidad,  $Pr(T > t|H_0)$ , que se denomina el *p-valor* de los datos.

En general, la hipótesis nula no comprende una única situación concreta sino un conjunto de ellas por lo que estas probabilidades calculadas en el supuesto de que sea cierta la hipótesis nula deben calcularse en aquella situación de las comprendidas dentro de la hipótesis nula más cercana a la alternativa, en la que esta probabilidad que pretendemos hacer tan pequeña como sea posible es máxima para garantizar que la probabilidad de cometer el error de

tipo I nunca es superior al calculado.

Para realizar un análisis estadístico sobre el rendimiento, se supone que disponemos de los resultados de la ejecución de  $k$  algoritmos heurísticos para  $n$  instancias. Sea  $c_{ji}$  una medida del rendimiento del  $j$ -ésimo algoritmo sobre la  $i$ -ésima instancia. La tarea consiste en contrastar estadísticamente si, basado en los valores  $c_{ji}$ , se puede afirmar que el rendimiento de los algoritmos es significativamente diferente, y en caso de que el número de algoritmos sea mayor que dos, cuales son los algoritmos cuyo rendimiento es realmente diferente. De forma mucho más ambiciosa se trata incluso de proporcionar una medida o evaluación de esas diferencias. Sin embargo, estas afirmaciones no pueden realizarse de forma aislada sin indicadores de las garantías que aportan los datos de que sean ciertas o del riesgo de que alguna sea incorrecta.

Frente a los datos de evaluación del rendimiento de algoritmos en una serie de instancias se pueden plantear diversos contrastes de hipótesis para realizar comparaciones objetivas y fiables. Esta situación es similar a la que se suele presentar en la ingeniería y las ciencias experimentales, donde el uso de técnicas estadísticas es más corriente. Por ejemplo en las pruebas para validar el efecto de un nuevo fármaco, la resistencia un nuevo material, los beneficios de diversos tratamientos médicos, la calidad obtenida con varios procesos de producción alternativos, etc. Los contrastes paramétricos son más conocidos que los no paramétricos pero con la creciente divulgación de éstos (ver [10], [18]) la elección en cada circunstancia debe realizarse teniendo en cuenta, además de las características propias de los contrastes, lo que realmente miden los estadísticos y las garantías de que sean ciertas las suposiciones que se asumen sobre los experimentos.

El cálculo de promedios entre los indicadores del rendimiento en diferentes instancias implica una postura en las comparaciones que es, al menos, discutible; los rendimientos sobre instancias de distinta dificultad o magnitud no son necesariamente comensurables. Es poco frecuente que en las tablas con valores del objetivo o tiempos de cómputo se ofrezcan promedios; seguramente porque no se encuentra justificación suficiente para su uso, y por tanto rechazarían procedimientos estadísticos que se basen en ellos.

Los promedios están justificados si los algoritmos se ejecutan sobre conjuntos de instancias relacionados, provenientes de un mismo contexto de aplicación o generadas al azar con las mismas características o muy similares. En este caso está justificado promediar entre los indicadores del rendimiento para las instancias similares, como en la ejecución repetida sobre una misma instancia de algoritmos con alguna componente aleatoria.

Los promedios se ven también afectados por los casos patológicos (outliers); permiten compensar un

resultado excelente con un conjunto amplio de malos resultados o que un fracaso rotundo en un dominio prevalezca sobre buenos resultados sobre una mayoría de los demás dominios. Pueden existir circunstancias en las que este comportamiento sea deseable, mientras que generalmente se prefiere un algoritmo que funciona bien en cuantos más casos mejor, lo que hace indeseable los promedios (prevalece la eficacia sobre la efectividad).

#### IV. COMPARACIÓN DE DOS ALGORITMOS

Para empezar con la situación más sencilla, supongamos que se ejecutan dos algoritmos  $A$  y  $A'$  sobre un conjunto de  $n$  instancias y el rendimiento en la  $i$ -ésima instancia es evaluado por los indicadores  $c_{1i}$  y  $c_{2i}$ . Frente al test de la  $t$  de student basado en la diferencia de los promedios, se proponen otros dos tests no paramétricos: el test de los rangos con signos de Wilcoxon [19] basado en la ordenación de las diferencias y el test de los signos que sólo tienen en cuenta el número de las diferencias positivas y las negativas. Aunque estos tests son más débiles (menos potentes) miden diferencias entre los algoritmos desde otro punto de vista.

##### A. El test $t$ para casos emparejados

Una forma usual de contrastar si las diferencias entre dos variables (los indicadores del rendimiento de dos algoritmos) en una muestra de casos (las instancias) es atribuible al azar, es calcular el test  $t$  para datos emparejados, que chequea si la diferencia promedio es significativamente diferente de cero. La diferencia en el rendimiento entre los dos algoritmos en la  $i$ -ésima instancia es  $d_i = c_{2i} - c_{1i}$ . Si

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i \quad \text{y} \quad \sigma_d^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2,$$

el estadístico  $T = \bar{d}/\sigma_{\bar{d}}$ , se distribuye según la  $t$  de Student con  $n-1$  grados de libertad. Para contrastar si existe diferencia entre el rendimiento de los dos algoritmos se realiza un contraste bilateral y, fijado el nivel  $\alpha$  y encontrado en la tabla el valor crítico  $t_{n,\alpha/2}$ , se rechaza la hipótesis nula si  $|T| > t_{n,\alpha/2}$ . Para contrastar que, por ejemplo, el segundo algoritmo tiene mejor rendimiento que el primero se aplica un contraste unilateral y, a partir del valor crítico es  $t_{n,\alpha}$ , se rechaza la hipótesis nula si  $T > t_{n,\alpha}$ .

Un primer inconveniente importante es que el test  $t$  requiere que las diferencias sobre las distintas instancias sean commensurables; por tanto, es tan cuestionable como el cálculo de promedios entre instancias. Además, a menos que el tamaño muestral sea suficientemente grande (30 instancias), se requiere también que la diferencia de las dos variables comparadas tenga distribución normal. No existe argumento suficiente a favor de esto en nuestro contexto y los tests de normalidad existentes, como el de Kolmogorov-Smirnov, son poco potentes para muestras pequeñas; es decir son incapaces de encontrar

anormalidades. Otro inconveniente es que el test  $t$  se ve gravemente afectado por los casos patológicos que sesgan el test quitándole potencia.

##### B. El Test de los rangos con signo de Wilcoxon

El test de los rangos con signos de Wilcoxon es la alternativa no paramétrica al test de la  $t$  de Student para casos emparejados. Para aplicar el test se calculan los valores absolutos de las diferencias en el rendimiento de dos algoritmos para cada instancia y se ordenan de mayor a menor y se comparan los lugares que ocupan las diferencias a favor de uno y otro algoritmo. Formalmente, sea  $rang(d_i)$  la posición que ocupa  $|d_i|$  en esta ordenación, denominado rango de  $d_i$ . Sea  $R^+$  y  $R^-$  las sumas respectivas de los rangos de las diferencias positivas y negativas. Los rangos de las diferencias nulas ( $d_i = 0$ ) se reparten entre ambas sumas; si hay un número impar de ellas, se ignora una. Por tanto:

$$R^+ = \sum_{d_i > 0} rang(d_i) + \frac{1}{2} \sum_{d_i = 0} rang(d_i)$$

$$R^- = \sum_{d_i < 0} rang(d_i) + \frac{1}{2} \sum_{d_i = 0} rang(d_i)$$

Sea  $T = \min(R^+, R^-)$  la menor de las sumas. Si la hipótesis nula es cierta  $T$  debe aproximarse a  $n(n+1)/4$  y cuanto más diferente sea el rendimiento de ambos algoritmos, menor se será  $T$ . Muchos textos de estadística incluyen valores críticos exactos para  $T$  para  $n$  hasta 25. Para un número mayor de instancias, el estadístico

$$z = \frac{T - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$$

tiene distribución aproximadamente normal. Por tanto, para  $\alpha = 0,05$ , la hipótesis nula del contraste bilateral se puede rechazar si  $z$  es menor  $-1,96$ .

TABLA I  
CÁLCULO DE LOS RANGOS.

| $i$ | $A$ | $A'$ | $d_i$ | $rang(d_i)$ |
|-----|-----|------|-------|-------------|
| 1   | 972 | 981  | +9    | 8           |
| 2   | 957 | 978  | +21   | 10          |
| 3   | 954 | 971  | +17   | 9           |
| 4   | 940 | 962  | +22   | 11          |
| 5   | 936 | 931  | -5    | 3,5         |
| 6   | 882 | 888  | +6    | 5           |
| 7   | 775 | 838  | +63   | 14          |
| 8   | 763 | 768  | +5    | 3,5         |
| 9   | 678 | 678  | 0     | 1,5         |
| 10  | 661 | 668  | +7    | 6           |
| 11  | 628 | 661  | +33   | 12          |
| 12  | 619 | 666  | +47   | 13          |
| 13  | 599 | 591  | -8    | 7           |
| 14  | 583 | 583  | 0     | 1,5         |

Ejemplo. Para ilustrar la aplicación de los test no paramétricos propuestos se usan datos supuestos muy similares a los expuestos en [4]. Supongamos que los datos de la tabla I corresponden a la comparación entre un algoritmo  $A$  y su mejora  $A'$ . Ambos algoritmos se ejecutaron sobre 14 instancias; se muestran las diferencias y el rango cada una de ellas. A las diferencias de igual magnitud se les asigna el rango promedio entre las que le corresponderían.

La suma de los rangos de las diferencias positivas es  $R^+ = 93$  y la de las negativas es  $R^- = 12$ , por tanto, el valor práctico del estadístico es  $T = 12$ . De la tabla de valores críticos exactos para el test de Wilcoxon con  $n = 14$  conjuntos y nivel de confianza  $\alpha = 0,05$ , se obtienen los valores críticos 21 para el contraste bilateral y 25 para el lateral. Por tanto, en ambos casos se puede rechazar la hipótesis nula y podemos aceptar que la diferencia entre los algoritmos  $A$  y  $A'$  es significativa y que el algoritmo  $A'$  mejora significativamente al algoritmo  $A$ .

El test de los rangos con signo de Wilcoxon es más sensible que el test  $t$ , requiere conmensurabilidad de las diferencias, pero sólo cualitativamente: mayores diferencias cuentan más, lo que puede ser deseable, pero se ignoran las magnitudes absolutas. Desde el punto de vista estadístico es más seguro porque no se asume distribución normal. Los *outliers* tienen menos efecto que en el test de la  $t$ . El test de Wilcoxon contempla diferencias continuas, por tanto los índices de rendimiento no deben ser redondeados ya que ello disminuiría la potencia del test debido al aumento artificial del número de empates. Cuando se dan las condiciones del test de la  $t$ , el test de Wilcoxon es menos potente, pero cuando se incumplen puede ser más potente que el test  $t$ .

### C. El test de los signos

Una forma sencilla y usual de comparar dos algoritmos es contar el número de veces en las que uno de ellos tiene mejor comportamiento que el otro. Bajo la hipótesis nula de que ambos tienen el mismo rendimiento, este número se aproximaría a la mitad de las veces, más concretamente, obedecería a una distribución binomial con probabilidad de éxito 0.5 y número de pruebas  $n$ . Los valores críticos de esta distribución se suelen encontrar tabulados en cualquier libro de estadística. Cuando  $n$  es suficientemente grande, la distribución es aproximadamente normal con media  $n/2$  y varianza  $n/4$ . Por tanto, al nivel  $\alpha$ , la hipótesis nula se puede rechazar, aceptando que  $A'$  es una mejora de  $A$  si el número de veces que  $A'$  gana al algoritmo  $A$  es al menos  $n/2 + z_\alpha\sqrt{n}/2$ . Para  $\alpha = 0,05$ , es  $z_\alpha = 1,96$ , lo que ha dado origen a la regla rápida de  $n/2 + \sqrt{n}$ .

Ejemplo. De la tabla I, el número de victorias del algoritmo  $A'$  es 10, con dos empates. Los empates se pueden repartir aunque lo más conservador es no contarlos, en favor de la hipótesis nula. Por tanto,

repartiendo los empates se puede concluir que el algoritmo  $A'$  es mejor que el  $A$  sólo al 90% de confianza pero con la actitud conservadora no se puede rechazar la hipótesis nula a este nivel. El correspondiente  $p$ -valor se obtiene buscando el valor de  $p$  tal que  $n/2 + z_p\sqrt{n}/2 = 10$  con la tabla de la normal.

Este test no requiere ninguna conmensurabilidad entre las diferencias ni normalidad en la distribución y por tanto es aplicable a cualquier conjunto de instancias. Es más débil que el test de Wilcoxon. Siguiendo el test de los signos se aceptaría la superioridad de la mejora frente al algoritmo original si tiene mejor rendimiento en una mayoría significativa de casos. Algunos autores argumentan además que sólo se deben tener en cuenta las diferencias significativas, y el umbral de significación se fija por algún otro criterio; el contraste es similar y se puede aplicar un test del mismo tipo.

## V. COMPARACIONES MÚLTIPLES

Los tests de comparaciones por pares no han sido diseñados para obtener conclusiones sobre varias variables. Para un estudio experimental con 7 algoritmos, no parece razonable realizar las 21 comparaciones por pares posibles. Cuando se realiza un alto número de contrastes, aunque todas las hipótesis nulas sean ciertas y el nivel de confianza en cada uno de ellos sea razonable, la probabilidad de que alguna sea rechazada por efectos del azar crece, siendo relativamente fácil llegar a la conclusión de que existen diferencias, sin haberlas.

La realización simultánea de múltiples comparaciones es un tema importante en el contraste de hipótesis estadística [8], [16]. Aunque recientemente se tiende también a controlar la tasa de descubrimientos falsos (porcentaje de hipótesis nulas erróneamente rechazadas), el objetivo tradicional es controlar el error global, la probabilidad de cometer al menos un error de tipo I en todas las comparaciones. La corrección de Bonferroni, consistente en dividir el nivel  $\alpha$  por el número de contrastes, es conservativa y débil ya que asume la independencia de las hipótesis. Existen procedimientos estadísticos más especializados como el bien conocido Análisis de la Varianza (ANOVA) y la alternativa no paramétrica; el test de Friedman. Este último, y especialmente su correspondiente test a posteriori de Nemenyi es quizás menos conocido e infrecuente, aunque más recomendable. Además, para evitar la redundancia en las conclusiones se utilizan procedimientos, como los procedimientos de Holm y Hochberg, para organizar el orden en el que se realizan los contrastes.

### A. ANOVA

El método corriente para contrastar la existencia de diferencias entre varias variables en un mismo conjunto de casos es el ANOVA de bloques [6]. En nuestro contexto, la hipótesis nula a contrastar es que todos los algoritmos rinden igual y que las difer-

encias observadas entre los rendimientos mostrados por los algoritmos son debidas al azar. El ANOVA divide la variabilidad total del indicador del rendimiento en tres sumandos: la variabilidad entre algoritmos (tratamientos en la terminología usual en ANOVA), la variabilidad entre instancias (bloques en la terminología ANOVA) y la variabilidad residual (o de error). Si la variabilidad entre algoritmos es significativamente mayor que la variabilidad residual, podemos rechazar la hipótesis nula y concluir que hay alguna diferencia entre los algoritmos. Este contraste se realiza con la distribución  $F$  de Fisher-Snedecor y en caso de rechazar la igualdad en el rendimiento de los algoritmos, se puede proceder con un contraste a posteriori para encontrar entre que pares de algoritmos se presentan esas diferencias. Entre los muchos tests para realizar esto en ANOVA, los dos más corrientes son el test de Tuckey para comparar todos los algoritmos entre si y el test de Dunnett para compararlos con un algoritmo control (por ejemplo, una heurística básica y algunas mejoras propuestas, o comparar los algoritmos novedosos propuestos con métodos existentes). Ambos métodos usan la estimación de la desviación típica de la diferencia entre los rendimientos de dos algoritmos dividiendo la variabilidad residual por el número de instancias menos uno. Para comparar pares de algoritmos, las correspondientes diferencias en rendimiento se dividen por esta estimación y se comparan con el valor crítico.

Los tests usuales en ANOVA están basados en suposiciones que son previsiblemente violadas cuando se analiza el rendimiento de algoritmos heurísticos de optimización. Tales requerimientos del ANOVA se refieren a la normalidad de las distribuciones, la igualdad de varianzas y la independencia de las medidas. La violación de estas suposiciones perjudica aún más a los contrastes a posteriori.

### B. Test de Friedman

El test de Friedman [7] es un test no paramétrico correspondiente al test  $F$  en un ANOVA de bloques. Se ordenan los  $k$  algoritmos separadamente para cada instancia según el rendimiento alcanzado, al mejor algoritmo tiene rango 1, el segundo mejor rango 2 y así sucesivamente. En caso de empate se asignan rangos intermedios. Sea  $r_j^i$  el rango del  $j$ -ésimo algoritmo sobre la  $i$ -ésima instancia. El test de Friedman compara los rangos medios de los algoritmos calculados por  $R_j = \frac{1}{n} \sum_{i=1}^n r_j^i$ .

Bajo la hipótesis nula, el rendimiento de todos los algoritmos es el mismo y los rangos  $R_j$  deben ser similares. El estadístico

$$\chi_F^2 = \frac{12n}{k(k+1)} \left[ \sum_{j=1}^k R_j^2 - \frac{k(k+1)^2}{4} \right]$$

propuesto por Friedman, se distribuye de acuerdo a

una  $\chi^2$  con  $k-1$  grados de libertad, cuando  $n$  y  $k$  son suficientemente grandes (como regla experimental práctica,  $n > 10$  y  $k > 5$ ). Para un número menor de algoritmos y casos se han calculado los valores críticos exactos [17], [20]. Pero en [12] se propone un estadístico mejor,  $F_F = \frac{(n-1)\chi_F^2}{n(k-1)-\chi_F^2}$  que se distribuye de acuerdo a la distribución  $F$  de Fisher-Snedecor con  $k-1$  y  $(k-1)(n-1)$  grados de libertad. La tabla con los valores críticos puede encontrarse en cualquier libro de estadística. Como en las comparaciones por pares, el test no paramétrico de Friedman tiene teóricamente menos potencia que el test  $F$ , cuando las suposiciones de ANOVA se verifican, pero no ocurre así en caso contrario.

### C. Contrastes a posteriori

Si se rechaza la hipótesis nula de equivalencia de los rendimientos de los algoritmos, se puede proceder a realizar contrastes a posteriori. El test de Nemenyi [15] es similar al de Tukey para ANOVA y se usa para comparar todos los algoritmos entre si. El rendimiento de dos algoritmos es significativamente diferente si los rangos promedios difieren, al menos en el valor crítico  $CD = q_\alpha \sqrt{\frac{k(k+1)}{6n}}$  donde los valores críticos  $q_\alpha$  son los del estadístico de rangos estudentizados dividido por  $\sqrt{2}$ .

Para comparar los algoritmos con un algoritmo de control se puede usar los métodos generales para controlar los errores comunes como la corrección de Bonferroni u otra similar que resulte más potente. El estadístico para comparar los rendimientos de los algoritmos  $i$ -ésimo y  $j$ -ésimo es:  $Z = \frac{R_i - R_j}{\sqrt{\frac{k(k+1)}{6n}}}$ .

A partir del valor práctico  $z$  se encuentra el correspondiente  $p$ -valor usando la distribución normal, que es comparado con el nivel apropiado. Los tests se diferencian en la forma en que ajustan el valor de  $\alpha$  para compensar las comparaciones múltiples.

El test de Bonferroni-Dunn [5] divide el valor de  $\alpha$  por el número de comparaciones realizadas. El método alternativo para realizar los mismos contrastes es calcular la diferencia crítica usando la misma fórmula que para el test de Nemenyi, pero usando los valores críticos para  $\alpha/(k-1)$ . La comparación entre las tablas de los test de Nemenyi y de Dunn muestran que la potencia del test a posteriori es mucho mayor cuando todos los algoritmos se comparan sólo con uno de control y no entre ellos. Por tanto no se deben hacer todas las comparaciones por pares cuando sólo se trata de ver si un método novedoso propuesto es mejor que los existentes.

A diferencia del procedimiento de Bonferroni-Dunn que hace los contrastes de una vez, los procedimientos secuenciales hacia arriba (*step-up*) y hacia abajo (*step-down*) contrastan las hipótesis ordenadamente según el nivel. Denotaremos los  $p$ -valores ordenados por  $p_1, p_2, \dots$ , de forma que  $p_1 \leq p_2 \leq \dots \leq p_{k-1}$ . Dos métodos simples usuales son el

método descendente de Holm [11] y el ascendente de Hochberg [9]. Ambos comparan cada  $p_i$  con  $\alpha/(k-i)$ , pero difieren en el orden de los tests. El procedimiento del test hacia abajo de Holm empieza con el  $p$ -valor más significativo. Si  $p_1$  está por debajo de  $\alpha/(k-1)$ , se rechaza la correspondiente hipótesis nula y se procede a comparar  $p_2$  con  $\alpha/(k-2)$ . Si se rechaza la segunda hipótesis se procede con la tercera, y así sucesivamente. Cuando una hipótesis nula no pueda ser rechazada, no se rechaza ninguna de las restantes. El procedimiento del test hacia arriba de Hochberg trabaja en la dirección contraria, comparando el mayor  $p$ -valor con  $\alpha$ , el siguiente más largo con  $\alpha/2$  y así hasta que encuentre una hipótesis nula que pueda rechazar. Entonces todas las hipótesis con  $p$ -valor menor se rechazan también.

El procedimiento de Holm es más potente que el de Bonferroni-Dunn y no hace ninguna suposición adicional sobre las hipótesis contrastadas. La única ventaja del test de Bonferroni-Dunn parece ser que es más fácil de describir y visualizar porque usa la misma diferencia crítica para todas las comparaciones. El método de Hochberg rechaza más hipótesis que el de Holm. Aunque aquí se describen los procedimientos de Holm y Hochberg como métodos para realizar contrastes a posteriori para el contraste de Friedman, pueden usarse cuando se contrastan a la vez múltiples hipótesis posiblemente de varios tipos.

Algunas veces el test de Friedman encuentra unas diferencias significativas pero los test a posteriori no llegan a detectarla. Esto se debe a la menor potencia de estos últimos. En este caso no se puede sacar ninguna otra conclusión de que el comportamiento de algunos algoritmos es diferente. En los experimentos, esto sólo ocurre en unos pocos casos entre miles.

Ejemplo. Consideremos los datos de la tabla II, que supuestamente compara un algoritmo heurístico  $A$  y con dos mejoras,  $A1$  y  $A2$ , obtenidas cada una por la incorporación de una nueva componente, y con la combinación de ambas mejoras (algoritmo  $A12$ ). Los rangos promedios proporcionan ya una idea intuitiva: los algoritmos  $A1$  y  $A12$  tienen rangos próximos a 2 y los algoritmos  $A$  y  $A2$  a 3.

El test de Friedman rechaza la igualdad si los rangos medios medidos son significativamente diferentes del rango medio esperado bajo la hipótesis nula para todos los algoritmos;  $R = 2,5$ . En este caso:  $\chi_F^2 = 8,85$  y  $F_F = 3,47$ . Con 4 algoritmos y 14 instancias,  $F_F$  se distribuye de acuerdo a la distribución  $F$  con 3 y 39 grados de libertad. El valor crítico de  $F(3, 39)$  al nivel  $\alpha = 0,05$  es 2,85 por lo que se rechaza la hipótesis nula.

Los análisis a posteriores dependen de lo que se pretenda estudiar. En ellos interviene la estimación del error estándar  $SE = \sqrt{\frac{4 \cdot 5}{6 \cdot 14}} = 0,49$ . Si no se distingue individualmente ningún algoritmo, se usa el test de Nemenyi para las comparaciones por pares.

El valor crítico es  $q_\alpha = 2,57$  y la correspondiente diferencia crítica es  $CD = 2,57 \cdot SE = 1,25$ . Puesto que incluso la diferencia entre el mejor y el peor rendimiento de los algoritmos es ya menor que este valor se puede concluir que el test a posteriori no es suficientemente potente para detectar ninguna diferencia significativa entre los algoritmos. Al nivel  $\alpha = 0,10$  la diferencia crítica es  $CD = 2,29 \cdot SE = 1,12$ .

Se pueden identificar dos grupos de algoritmos: el rendimiento del algoritmo  $A$  es significativamente peor que las mejoras  $A1$  y  $A12$ , pero el algoritmo  $A2$  no se puede encuadrar claramente en ninguno de estos grupos. La conclusión estadística correcta sería que los datos experimentales no son suficientes para alcanzar una conclusión con respecto al algoritmo  $A2$ . La otra hipótesis posible hecha antes de recoger los datos sería que es posible mejorar el rendimiento del algoritmo  $A$  ajustando algún parámetro. La forma más fácil de verificar esto es calcular la diferencia crítica con el test de Bonferroni-Dunn. En la tabla correspondiente encontramos que el valor crítico  $q_{0,05}$  para 4 algoritmos es 2,39, y la diferencia crítica es  $CD = 2,39 \cdot SE = 1,17$ .

El rendimiento del algoritmo  $A12$  es significativamente mejor que el algoritmo  $A$  ( $3,14 - 1,96 > 1,17$ ) pero  $A2$  no es significativamente mejor que el algoritmo  $A$  ( $3,14 - 2,89 < 1,17$ ), mientras que el algoritmo  $A1$  está debajo de la diferencia crítica, aunque cerca de ella ( $3,14 - 2,00 < 1,17$ ). Se puede concluir que los experimentos muestran que la mejora "1" parece que ayuda mientras que no se detecta ninguna mejora significativa con la mejora "2".

Otro planteamiento factible es pretender que las modificaciones propuestas mejoran el algoritmo inicial  $A$ . Para ello se calculan y ordenan los estadísticos y  $p$  valores correspondientes como aparece en la tabla III. El procedimiento de Holm rechaza la primera hipótesis y entonces también la segunda porque los correspondientes  $p$ -valores son menores que los valores de  $\alpha$  ajustados. La tercera hipótesis no puede ser rechazada, si hubiera más hipótesis tampoco se rechazarían. El procedimiento de Hochberg empieza por abajo. Incapaz de rechazar la última hipótesis, comprueba la penúltima, la rechaza y con ella todas las hipótesis con  $p$ -valores menores (mayores diferencias). Los procedimientos hacia abajo y hacia arriba encuentran que  $A12$  y  $A1$  son significativamente diferentes de  $A$ , mientras que el test de Bonferroni-Dunn test encuentra que los algoritmos  $A$  y  $A2$  son también similares.

El método ANOVA y el test de Friedman pueden considerar observaciones múltiples por celda, supuesto que las observaciones son independientes. Este situación se presenta cuando se usan conjuntos de instancias de las mismas características o ejecución repetida sobre una misma instancia; los promedios correspondientes se usan como indicadores del rendimiento en cada conjunto.

TABLA II  
COMPARACIÓN DE CUATRO ALGORITMOS.

| $i$ | A    |     | A1   |     | A2   |     | A12  |     |
|-----|------|-----|------|-----|------|-----|------|-----|
| 1   | 972  | 4   | 981  | 1   | 975  | 2   | 975  | 3   |
| 2   | 957  | 3   | 978  | 1   | 946  | 4   | 970  | 2   |
| 3   | 954  | 4   | 971  | 1   | 968  | 2   | 967  | 3   |
| 4   | 940  | 4   | 962  | 2,5 | 965  | 1   | 962  | 2,5 |
| 5   | 936  | 1   | 931  | 2,5 | 916  | 4   | 931  | 2,5 |
| 6   | 882  | 4   | 888  | 2   | 886  | 3   | 898  | 1   |
| 7   | 775  | 4   | 838  | 3   | 866  | 2   | 875  | 1   |
| 8   | 763  | 4   | 768  | 3   | 771  | 2   | 798  | 1   |
| 9   | 678  | 2,5 | 678  | 2,5 | 678  | 2,5 | 678  | 2,5 |
| 10  | 661  | 3   | 668  | 2   | 609  | 4   | 685  | 1   |
| 11  | 628  | 4   | 661  | 1   | 654  | 3   | 657  | 2   |
| 12  | 619  | 3   | 666  | 2   | 614  | 4   | 669  | 1   |
| 13  | 599  | 1   | 591  | 2   | 590  | 3   | 569  | 4   |
| 14  | 583  | 2,5 | 583  | 2,5 | 563  | 4   | 625  | 1   |
|     | 3,14 |     | 2,00 |     | 2,89 |     | 1,96 |     |

TABLA III  
CONTRASTES A POSTERIORI.

|   | Alg | $Z = (R_0 - R_i)/SE$              | $p$   | $\alpha/i$ |
|---|-----|-----------------------------------|-------|------------|
| 1 | A2  | $\frac{3,14 - 2,89}{0,49} = 0,51$ | 0,607 | 0,050      |
| 2 | A1  | $\frac{3,14 - 2,00}{0,49} = 2,33$ | 0,019 | 0,025      |
| 3 | A12 | $\frac{3,14 - 1,96}{0,49} = 2,41$ | 0,016 | 0,017      |

## VI. CONCLUSIONES

Cada vez son menos las dificultades para aplicar técnicas estadísticas en el análisis de los resultados de los experimentos con heurísticas. En [3] se muestran como reevaluar en esta línea algunos experimentos aparecidos en la literatura. Un trabajo más extenso [1] analiza de forma más completa el uso del análisis estadístico en la comparación de heurísticas. En la disyuntiva entre optar por métodos paramétricos o no paramétricos hay que tener en cuenta que, como regla general, un test no paramétrico tiene teóricamente menos potencia que el correspondiente test paramétrico, cuando las suposiciones de éste se verifican, pero no ocurre así en caso contrario [10], [18]. A pesar de todo, las conclusiones con ambas metodologías suele ser concordantes [17]. Estos test no paramétrico son explícitamente en [4] para un contexto muy cercano y están siendo cada vez más relevante en Genómica [13] donde se precisa realizar una gran cantidad de contrastes de forma simultánea.

## AGRADECIMIENTOS

Este trabajo ha sido parcialmente sbvencionado por los proyectos TIN2005-08404-C04-03 (70 % FEDER) y PI042005/044.

## REFERENCIAS

- [1] Adenso-Díaz, B., Corral N. A guide to statistical analysis for heuristics evaluation. Internal Technical Report. Universidad de Oviedo, 2003.
- [2] Barr, R.S., Golden, B.L., Kelly, J.P. Resende, M.G.C., Stewart, W.R. Design and Reporting on Computational Experiments with heuristics methods. *Journal of Heuristics* 1(1):9-32, 1995
- [3] Coffin, M., Saltzman, M.J. Statistical Analysis of Computational Tests of Algorithms and Heuristics. *INFORMS Journal on Computing* Vol. 12(1), 24-44 (2000)
- [4] J. Demšar, Statistical Comparisons of Classifiers over Multiple Data Sets, *Journal of Machine Learning Research*, 7:1-30, 2006.
- [5] Dunn, O. J. Multiple comparisons among means. *Journal of the American Statistical Association*, 56:52-64, 1961.
- [6] Fisher, R.A.. *Statistical methods and scientific inference*. Hafner Publishing Co., 1959.
- [7] Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32:675-701, 1937.
- [8] Hochberg, Y. and Tamhane, A. C. *Multiple Comparison Procedures*. Wiley, 1987.
- [9] Hochberg, Y. 1988. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75: 800-802
- [10] Hollander, M. and Wolfe, D. *Nonparametric Statistical Methods*. Wiley, 1999
- [11] Holm, S. 1979. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 6: 65-70.
- [12] Iman, R. L. Davenport, J. M. Approximations of the critical region of the Friedman statistic. *Communications in Statistics*, 571-595, 1980.
- [13] K.F. Manly, D. Nettleton, J.T. Gene Hwang. Genomics, Prior Probability, and Statistical Tests of Multiple Hypotheses. *Genome Research* 14:997-1001, 2004
- [14] Melián Batista, B., Moreno Pérez, J.A. Moreno Vega, J.M. Metaheurísticas: una vision global. *Inteligencia Artificial*. Número 19. Vol. 2, 7-28. 2003.
- [15] Nemenyi, P. B. *Distribution-free multiple comparisons*. PhD thesis, Princeton University, 1963.
- [16] Shaffer, J. P. Multiple hypothesis testing. *Annual Review of Psychology*, 46:561-584, 1995.
- [17] Sheskin, D. J. *Handbook of parametric and nonparametric statistical procedures*. Chapman and Hall, 2000.
- [18] Sprent, P. Smeeton, N.C. *Applied Nonparametric Statistical Methods*. Chapman and Hall, 2001.
- [19] Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics*, 1:80-83, 1945.
- [20] Zar, J. H. *Biostatistical Analysis*. Prentice Hall, 1998.