Un Tutorial Metodológico para hacer Comparaciones Estadísticas con Tests No Paramétricos en Propuestas de Minería de Datos

Salvador García Joaquín Derrac Francisco Herrera
Dept. Informática Dept. CCIA - CITIC Dept. CCIA - CITIC
Univ. de Jaén Univ. de Granada Univ. de Granada
sglopez@ujaen.es jderrac@decsai.ugr.es herrera@decsai.ugr.es

Resumen

En minería de datos, es necesario realizar análisis experimentales que nos permitan comprobar cuándo una técnica se comporta mejor que otra en un determinado problema. Los tests estadísticos no paramétricos han demostrado ser un conjunto de técnicas realmente efectivas y utilizadas frecuentemente para hacer comparaciones en entornos multi-dominio (en los que se usan varias instancias del problema a tratar).

En este trabajo, presentamos una metodología sobre el uso de tests no paramétricos basada en las necesidades del investigador en minería de datos cuando propone un nuevo método. Los tests más avanzados serán presentados como alternativa fiable a los clásicos y utilizaremos un caso de estudio en clasificación con varios conjuntos de datos para ejemplificar su uso.

1. Introducción

El análisis experimental del rendimiento de un método en Minería de Datos (MD) es una tarea crucial para determinar la veracidad de una conclusión obtenida. Decidir cuándo un algoritmo es mejor que otro no es una tarea trivial y depende de muchos factores. La inferencia estadística es una herramienta por la cual se puede estimar, estableciendo primero un llamado nivel de significancia, cuando un conjunto finito de resultados obtenido por un método es diferente de otro [7].

Los tests estadísticos pueden ser paramétricos y no paramétricos dependiendo del tipo de datos con el que trabajan [7]. Los primeros trabajan con valores númericos y son sensibles a *outliers* y suposiciones de independencia, normalidad y homoscedasticidad en la muestra de resultados, mientras que los segundos operan con valores ordinales (rankings) y son más flexibles que los primeros en el tipo de resultados donde pueden ser aplicados. Por esta razón, los tests no paramétricos están siendo muy eficaces y en entornos multi-dominio, donde se evaluán varias instancias del problema de MD a tratar, como es el procedimiento habitualmente [4, 6]. Por ejemplo, en clasificación, un entorno multi-dominio supone la evaluación de algoritmos considerando múltiples conjuntos de datos.

Centrándonos en tests no paramétricos, la posibilidad de elección de uno u otro dependerá de las necesidades del usuario así como de las características de los resultados que se van a analizar [3, 2]. En este trabajo proporcionamos una ayuda para resolver dicha cuestión cuando un investigador se propone comparar su propuesta con un conjunto de métodos. En primer lugar, enumeramos todas las técnicas estadísticas no paramétricas que se pueden utilizar en este tipo de comparaciones. Junto a esto, proporcionamos pautas acerca del uso de dichas técnicas y damos ejemplos prácticos de uso. Finalmente, resolvemos preguntas frecuentes sobre su uso.

2. Preliminares: Marco Experimental

Nos centramos en el problema de clasificación y utilizamos 24 datasets recogidos del repositorio UCI y KEEL dataset [1]. Los algoritmos que utilizamos están integrados en la plataforma software KEEL [1] y son PDFC, NNEP, IS-CHC+1NN y FH-GBML. Los valores de los parámetros de los algoritmos considerados son los estándar recogidos en KEEL. Usamos la validación cruzada en 10 *folds* y los resultados medios obtenidos corresponden al procentaje de acierto en test sobre 3 ejecuciones de cada algoritmo. Los resultados obtenidos podrán visualizarse en posteriores secciones del trabajo donde explicamos el funcionamiento de los tests estadísticos.

3. Técnicas Clásicas para Comparar una Propuesta con Varias

Se revisarán dos tests estadísticos para el fin descrito en este trabajo: el test de Friedman y el test de signos múltiple.

3.1. Test de Friedman

Sea r_i^j el ranking del j-ésimo de k algoritmos sobre el i-ésimo de n conjuntos de datos. El test de Friedman requiere el cálculo de los rankings medios de los algoritmos, $R_j = \frac{1}{n}\sum_i r_i^j$. Bajo la hipótesis nula que indica que todos los algoritmos se comportan similarmente, por lo que sus rankings R_j deben ser iguales, el estadístico de Friedman

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

se distribuye de acuerdo a una χ_F^2 con $k-1$ grados de libertad.

Iman y Davenport mostraron que el estadístico de Friedman presenta un comportamiento conservativo y propusieron un estadístico mejor

$$F_F = \frac{(n-1)\chi_F^2}{n(k-1)\chi_F^2}$$

que se distribuye de acuerdo a una distribución F con $k-1$ y $(k-1)(n-1)$ grados de libertad. Ver la Tabla A10 en [7] para encontrar los valores críticos. Si se rechaza la hipótesis nula, podemos proceder con un test a posteriori que se detallarán en la Sección 5.

Tabla 1: Ejemplo de uso del test de Friedman. Los rankings en paréntesis se usan en el cálculo del estadístico

Conjunto	PDFC	NNEP	IS-CHC+1NN	FH-GBML
adult	0,752 (4)	0,773 (3)	0,785 (2)	0,795 (1)
breast	0,727 (2)	0,748 (1)	0,724 (3)	0,713 (4)
bupa	0,736 (1)	0,716 (2)	0,585 (4)	0,638 (3)
car	0,994 (1)	0,861 (3)	0,880 (2)	0,791 (4)
cleveland	0,508 (4)	0,553 (2)	0,575 (1)	0,515 (3)
contraceptive	0,535 (2)	0,536 (1)	0,513 (3)	0,471 (4)
dermatology	0,967 (1)	0,871 (3)	0,954 (2)	0,532 (4)
ecoli	0,831 (1)	0,807 (3)	0,819 (2)	0,768 (4)
german	0,745 (1)	0,702 (4)	0,719 (2)	0,705 (3)
glass	0,709 (1)	0,572 (4)	0,669 (2)	0,607 (3)
haberman	0,722 (4)	0,728 (2)	0,725 (3)	0,732 (1)
iris	0,967 (1)	0,947 (4)	0,953 (3)	0,960 (2)
lymphography	0,832 (1)	0,752 (3)	0,802 (2)	0,691 (4)
mushrooms	0,998 (1)	0,992 (2)	0,482 (4)	0,910 (3)
newthyroid	0,963 (1,5)	0,963 (1,5)	0,954 (3)	0,926 (4)
penbased	0,982 (1)	0,953 (2)	0,932 (3)	0,630 (4)
ring	0,978 (1)	0,773 (4)	0,834 (3)	0,849 (2)
satimage	0,854 (1)	0,787 (3)	0,841 (2)	0,779 (4)
shuttle	0,965 (3)	0,984 (2)	0,995 (1)	0,947 (4)
spambase	0,924 (1)	0,887 (2)	0,861 (3)	0,804 (4)
thyroid	0,929 (3)	0,942 (1)	0,931 (2)	0,921 (4)
vehicle	0,837 (1)	0,643 (2)	0,602 (3)	0,554 (4)
wine	0,972 (1)	0,956 (2)	0,944 (3)	0,922 (4)
wisconsin	0,958 (4)	0,959 (3)	0,964 (1,5)	0,964 (1,5)
ranking medio	1,771	2,479	2,479	3,271

El procedimiento se ilustra en la Tabla 1, que compara los cuatro algoritmos considerados en el marco experimental. Los rankings medios proporcionan una comparación interesante de los algoritmos. En media, PDFC obtuvo el primer ranking 1,771; NNEP y IS-CHC+1NN obtuvieron el segundo y tercer ranking, con el mismo valor 2,479; y el último es FH-GBML con ranking 3,271. El test de Friedman prueba si los rankings medios obtenidos son significativamente diferentes de el ranking medio esperado bajo la hipótesis nula $R_j = 2{,}5$:

$$(Friedman)\ \chi_F^2 = \frac{12\cdot 24}{4\cdot 5}\cdot$$

$$\cdot\left[(1{,}771^2 + 2{,}479^2 + 2{,}479^2 + 3{,}271^2) - \frac{4\cdot 5^2}{4}\right] = 16{,}225$$

$$(Iman - Davenport)\ F_F = \frac{23\cdot 16{,}225}{24\cdot 3 - 16{,}225} = 6{,}691$$

Con cuatro algoritmos y 24 conjuntos, F_F se distribuye de acuerdo a una distribución F con $4-1=3$ y $(4-1)\cdot(24-1)=69$ grados de libertad. El valor p calculado usando la distribución $F(3,69)$ es $4{,}97\cdot 10^{-4}$, por lo que la hipótesis nula se rechaza con una alta probabilidad.

3.2. Test de Signos Múltiple

El siguiente procedimiento nos permite comparar directamente un conjunto de método con un método control etiquetado como 1. La técnica es una extensión del test de signos convencional [4], y lleva a cabo los siguientes pasos:

1. Representar con x_{i1} and x_{ij} los valores de rendimientos del método control y del j-ésimo en el conjunto i-ésimo.
2. Calcular las diferencias con signos $d_{ij} = x_{ij} - x_{i1}$. En otras palabras, emparejar cada rendimiento con el control y, en cada conjunto de datos, sustraer el rendimiento control del j-ésimo método.
3. Sea r_j igual al número de diferencias, d_{ij}, que tienen el signo menos frecuente (o positivo o negativo) dentro de un emparejamiento de un algoritmo con el control.
4. Sea M_1 la respuesta mediana de una muestra de resultados de la muestra control y M_j la respuestas mediana de una muestra de resultados del algoritmo j-ésimo. Aplicar una de las dos reglas de decisión siguientes:
   - Para comprobar $H_0 : M_j \geq M_1$ contra $H_1 : M_j < M_1$, rechaza H_0 si el número de signos más es menor o igual al valor crítico de R_j que aparece en la Tabla A.1 de [5] para $k-1$ (número de algoritmos excluyendo el control), n y la tasa de error experimental escogida.
   - Para comprobar $H_0 : M_j \leq M_1$ contra $H_1 : M_j > M_1$, rechaza H_0 si el número de signos menos es menor o igual que el valor crítico de R_j que aparece en la Tabla A.1 de [5] para $k-1$, n y la tasa de error experimental escogida.

La Tabla 2 muestra las cálculos realizados por este procedimiento.

Suponemos un nivel de significancia de $\alpha = 0{,}05$ y nuestras hipótesis son $H_0 : M_j \geq M_0$ y $H_1 : M_j < M_0$; esto es, nuestro algoritmo control PDFC es mejor que el resto de clasificadores. La referencia a la Tabla A.1 de [5] para $(k-1)=3$ y $n=24$ revela que el valor crítico de r_j es 6. Puesto que el número de signos más en la comparación en pareja entre el método control y IS-CHC+1NN y FH-GBML es menor que 6, entonces PDFC presenta mejor rendimiento que ellos. Sin embargo, la hipótesis nula no puede ser rechazada en la comparación entre PDFC y NNEP, por lo que conluimos que se comportan de forma similar.

Tabla 2: Comparación considerando tasa de acierto con PDFC como método control. Los signos en paréntesis se usan en el cálculo del Test de Signos Múltiple

	PDFC	NNEP	IS-CHC+1NN	FH-GBML
Conjunto	1 (Control)	2	3	4
adult	0,752	0,773 (+)	0,785 (+)	0,795 (+)
breast	0,727	0,748 (+)	0,724 (-)	0,713 (-)
bupa	0,736	0,716 (-)	0,585 (-)	0,638 (-)
car	0,994	0,861 (-)	0,880 (-)	0,791 (-)
cleveland	0,508	0,553 (-)	0,575 (+)	0,515 (+)
contraceptive	0,535	0,536 (+)	0,513 (-)	0,471 (-)
dermatology	0,967	0,871 (-)	0,954 (-)	0,532 (-)
ecoli	0,831	0,807 (-)	0,819 (-)	0,768 (-)
german	0,745	0,702 (-)	0,719 (-)	0,705 (-)
glass	0,709	0,572 (-)	0,669 (-)	0,607 (-)
haberman	0,722	0,728 (+)	0,725 (+)	0,732 (+)
iris	0,967	0,947 (-)	0,953 (-)	0,960 (-)
lymphography	0,832	0,752 (-)	0,802 (-)	0,691 (-)
mushrooms	0,998	0,992 (-)	0,482 (-)	0,910 (-)
newthyroid	0,963	0,963 (=)	0,954 (-)	0,926 (-)
penbased	0,982	0,953 (-)	0,932 (-)	0,630 (-)
ring	0,978	0,773 (-)	0,834 (-)	0,849 (-)
satimage	0,854	0,787 (-)	0,841 (-)	0,779 (-)
shuttle	0,965	0,984 (+)	0,995 (+)	0,947 (-)
spambase	0,924	0,887 (-)	0,861 (-)	0,804 (-)
thyroid	0,929	0,942 (+)	0,931 (+)	0,921 (-)
vehicle	0,837	0,643 (-)	0,602 (-)	0,554 (-)
wine	0,972	0,956 (-)	0,944 (-)	0,922 (-)
wisconsin	0,958	0,959 (+)	0,964 (+)	0,964 (+)
número de menos		16	18	20
número de mas		7	6	4
r_j		7	6	4

4. Técnicas Avanzadas para Comparar una Propuesta con Varias

Se presentan dos técnicas avanzadas para realizar comparaciones estadísticas múltiples con método control: el test de Friedman Alineado y el test de Quade.

4.1. Test de Friedman Alineado

El test de Friedman se basa en n conjuntos de rankings, un conjunto por cada relación de datos en nuestro caso; y los los rendimientos de los algoritmos analizados se ordenan separadamente para cada conjunto de datos. Dicho esquema permite hacer comparaciones intra-conjunto, puesto que las inter-conjunto no son significativas. Cuando el número de algoritmos en la comparación es pequeño, este procedimiento tiene cierta desventaja. En estos casos, la comparación entre conjuntos de datos es deseable y podemos aplicar el método de los rankings alineados.

En esta técnica, se calcula el rendimiento medio alcanzado por cada algoritmo en cada conjunto de datos, llamado valor de localización. Después, calcula las diferencias entre el rendimiento obtenido por cada algoritmo con respecto al valor de localización. Este paso se repite para todos los algoritos y conjuntos de datos. Las diferencias resultantes, llamadas observaciones alineadas, se ordenan desde 1 hasta kn de forma relativa unas con otras. Entonces, 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. Los rankings asignados a las observaciones alineadas se denominan rankings alineados.

El test de Friedman Alineado puede escribirse como

$$T = \frac{(k-1)\left[\sum_{j=1}^{k} \hat{R_{.j}}^2 - (kn^2/4)(kn+1)^2\right]}{\{[kn(kn+1)(2kn+1)]/6\} - (1/k)\sum_{i=1}^{n} \hat{R_{i.}}^2}$$

donde $\hat{R}_{i.}$ es igual al ranking total del i-ésimo conjunto de datos y $\hat{R}_{.j}$ es el ranking total del j-ésimo algoritmo.

El estadístico T se compara con una distribución chi cuadrado para $k-1$ grados de libertad. Los valores críticos pueden encontrarse en la Tabla A3 de [7]. Si la hipótesis nula es rechazada, podemos proceder con un test posterior de identificación de diferencias (ver Sección 5).

Ilustramos el test de Friedman Alineado con el experimento global asociado a este trabajo. La Tabla 3 muestra ls observaciones alineadas y los rankings alineados en paréntesis considerando el marco experimental expuesto en la Sección 2.

Tabla 3: Observaciones alineadas de los 4 algoritmos considerados. Los rankings en paréntesis son los usados en el cálculo del test de Friedman Alineado.

Conjunto	PDFC	NNEP	IS-CHC-1NN	FH-GBML
adult	-0,024 (74)	-0,003 (56)	0,009 (39)	0,019 (30)
breast	-0,001 (51)	0,020 (29)	-0,004 (59)	-0,015 (68)
bupa	0,068 (11)	0,047 (16)	-0,084 (90)	-0,031 (81)
car	0,112 (7)	-0,020 (72)	-0,002 (53)	-0,091 (92)
cleveland	-0,030 (80)	0,016 (32)	0,037 (19)	-0,023 (73)
contraceptive	0,022 (28)	0,022 (26)	-0,001 (50)	-0,043 (85)
dermatology	0,136 (4)	0,040 (17)	0,123 (5)	-0,299 (95)
ecoli	0,025 (24)	0,001 (48)	0,013 (33)	-0,038 (84)
german	0,027 (22)	-0,016 (69)	0,001 (47)	-0,013 (67)
glass	0,069 (10)	-0,068 (88)	0,030 (21)	-0,032 (82)
haberman	-0,005 (61)	0,002 (46)	-0,002 (54)	0,005 (41)
iris	0,010 (38)	-0,010 (66)	-0,003 (58)	0,003 (42)
lymphography	0,063 (13)	-0,017 (71)	0,032 (20)	-0,078 (89)
mushrooms	0,152 (2)	0,146 (3)	-0,363 (96)	0,065 (12)
newthyroid	0,012 (34,5)	0,012 (34,5)	0,002 (45)	-0,026 (76)
penbased	0,108 (8)	0,078 (9)	0,058 (14)	-0,244 (94)
ring	0,120 (6)	-0,085 (91)	-0,025 (75)	-0,010 (65)
satimage	0,038 (18)	-0,028 (79)	0,026 (23)	-0,036 (83)
shuttle	-0,008 (62)	0,012 (36)	0,022 (27)	-0,026 (77)
spambase	0,055 (15)	0,018 (31)	-0,008 (63)	-0,065 (87)
thyroid	-0,001 (52)	0,011 (37)	0,000 (49)	-0,010 (64)
vehicle	0,178 (1)	-0,016 (70)	-0,057 (86)	-0,105 (93)
wine	0,024 (25)	0,007 (40)	-0,004 (60)	-0,027 (78)
wisconsin	-0,003 (57)	-0,002 (55)	0,003 (43,5)	0,003 (43,5)
total	703,5	1121,5	1129,5	1701,5
ranking medio	29,313	46,729	47,063	70,896

De nuevo, los rankings medios proporcionan una buena comparación de los algoritmos. En media, PDFC es el mejor con ranking 29,313; NNEP y IS-CHC+1NN se sitúan segundos y terceros con rankings 46,729 y 47,063, respectivamente; y el último es FH-GBML con ranking 70,896. El test de Friedman Alineado comprueba si la suma de rankings alineados es significativamente diferente al ranking alineado total $\hat{R}_j = 1164$ esperado bajo la hipótesis nula:

$$\sum_{j=1}^{k} \hat{R_{.j}}^2 = 703{,}5^2+1121^2+1129{,}5^2+1701{,}5^2 = 5923547$$

$$\sum_{i=1}^{n} \hat{R_{i.}}^2 = 199^2+207^2+198^2+...+199^2 = 926830$$

$$T = \frac{(4-1)\left[5923547-(4\cdot 24^2/4)(4\cdot 24+1)^2\right]}{\{[4\cdot 24(4\cdot 24+1)(2\cdot 4\cdot 24+1)]/6\}-(1/4)\cdot 926830} =$$

$$= 18{,}837$$

Con 4 algoritmos y 24 conjuntos de datos, T se distribuye de acuerdo a una distribución chi cuadrado con $4-1=3$ grados de libertad. El valor p calculado usando la distribución $\chi^2(3)$ es $2{,}96\cdot 10^{-4}$, por lo que la hipótesis nula es rechazada con un alto nivel de significancia.

4.2. Test de Quade

El test de Friedman considera que todos los conjuntos de datos son iguales en importancia. Una alernativa a esto podría tener en cuenta que algunos conjuntos de datos son más difíciles o que las diferencias registradas en la ejecución de varios algoritmos sobre ellos son más distantes. Los rankings calculados en cada conjunto de datos podrían escalarse dependiendo de las diferencias observadas en los rendimientos de los algoritmos. El test de Quade lleva a cabo un análisis con rankings ponderados de las muestras de resultados.

El procedimiento comienza encontrandos los rankings r_i^j 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 conjuntos de datos de acuerdo al tamaño del rango de la muestra en cada conjunto. El rango de la muestra en un conjunto de datos i es la diferencia entre la observación más alta y baja en dicho conjunto de datos.

Rango en conjunto i: $i = \max_j\{x_{ij}\} - \min_j\{x_{ij}\}$.

Obviamente, hay n rangos muestrales, uno por cada conjunto de datos. 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 utiliza rankings medios en caso de empate. Sean $Q_1, Q_2, ..., Q_n$ los rankings asignados a los conjuntos de datos 1, 2, ..., n, respectivamente.

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

$$S_{ij} = Q_i\left[r_i^j - \frac{k+1}{2}\right]$$

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

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

$$W_{ij} = Q_i\left[r_i^j\right]$$

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 $j = 1, 2, ..., k$. Después, debemos calcular los términos:

$$A_2 = n(n+1)(2n+1)(k)(k+1)(k-1)/72$$

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

El estadístico es

$$T_3 = \frac{(n-1)B}{A_2 - B}$$

que está distribuido de acuerdo a una distribución F con $k-1$ y $(k-1)(n-1)$ grados de libetad. La tabla de valores críticos para esta distribución puede consultarse en [7], Tabla A10. Si la hipótesis nula es rechazada, podemos proceder con un test posterior de identificación de diferencias (ver Sección 5)

Tabla 4: Comparación de la tasa de acierto para los 4 algoritmos considerados en nuestro estudio. Los rankings en paréntesis se utilizan en el cálculo del test de Quade. S_{ij} y W_{ij} se muestran en este orden.

Conjunto	Muestra Rango	Ranking Q_i	PDFC	NNEP	IS-CHC-1NN	FH-GBML
adult	0,043	8	0,752 (12)(32)	0,773 (4)(24)	0,785 (-4)(16)	0,795 (-12)(8)
breast	0,035	5	0,727 (-2,5)(10)	0,748 (-7,5)(5)	0,724 (2,5)(15)	0,713 (7,5)(20)
bupa	0,151	18	0,736 (-27)(18)	0,716 (-9)(36)	0,585 (27)(72)	0,638 (9)(54)
car	0,203	19	0,994 (-28,5)(19)	0,861 (9,5)(57)	0,880 (-9,5)(38)	0,791 (28,5)(76)
cleveland	0,067	13	0,508 (19,5)(52)	0,553 (-6,5)(26)	0,575 (-19,5)(13)	0,515 (6,5)(39)
contraceptive	0,065	12	0,535 (-6)(24)	0,536 (-18)(12)	0,513 (6)(36)	0,471 (18)(48)
dermatology	0,436	23	0,967 (-34,5)(23)	0,871 (11,5)(69)	0,954 (-11,5)(46)	0,532 (34,5)(92)
ecoli	0,063	11	0,831 (-16,5)(11)	0,807 (5,5)(33)	0,819 (-5,5)(22)	0,768 (16,5)(44)
german	0,043	7	0,745 (-10,5)(7)	0,702 (10,5)(28)	0,719 (-3,5)(14)	0,705 (3,5)(21)
glass	0,137	16	0,709 (-24)(16)	0,572 (24)(64)	0,669 (-8)(32)	0,607 (8)(48)
haberman	0,010	2	0,722 (3)(8)	0,728 (-1)(4)	0,725 (1)(6)	0,732 (-3)(2)
iris	0,020	3	0,967 (-4,5)(3)	0,947 (4,5)(12)	0,953 (1,5)(9)	0,960 (-1,5)(6)
lymphography	0,141	17	0,832 (-25,5)(17)	0,752 (8,5)(51)	0,802 (-8,5)(34)	0,691 (25,5)(68)
mushrooms	0,515	24	0,998 (-36)(24)	0,992 (-12)(48)	0,482 (36)(96)	0,910 (12)(72)
newthyroid	0,038	6	0,963 (-6)(9)	0,963 (-6)(9)	0,954 (3)(18)	0,926 (9)(24)
penbased	0,352	22	0,982 (-33)(22)	0,953 (-11)(44)	0,932 (11)(66)	0,630 (33)(88)
ring	0,205	20	0,978 (-30)(20)	0,773 (30)(80)	0,834 (10)(60)	0,849 (-10)(40)
satimage	0,075	14	0,854 (-21)(14)	0,787 (7)(42)	0,841 (-7)(28)	0,779 (21)(56)
shuttle	0,048	9	0,965 (4,5)(27)	0,984 (-4,5)(18)	0,995 (-13,5)(9)	0,947 (13,5)(36)
spambase	0,120	15	0,924 (-22,5)(15)	0,887 (-7,5)(30)	0,861 (7,5)(45)	0,804 (22,5)(60)
thyroid	0,021	4	0,929 (2)(12)	0,942 (-6)(4)	0,931 (-2)(8)	0,921 (6)(16)
vehicle	0,282	21	0,837 (-31,5)(21)	0,643 (-10,5)(42)	0,602 (10,5)(63)	0,554 (31,5)(84)
wine	0,050	10	0,972 (-15)(10)	0,956 (-5)(20)	0,944 (5)(30)	0,922 (15)(40)
wisconsin	0,006	1	0,958 (1,5)(4)	0,959 (0,5)(3)	0,964 (-1)(1,5)	0,964 (-1)(1,5)
suma of rankings S_j			-332	11	27,5	293,5
rankings medios $T_j = \frac{W_j}{n(n+1)/2}$			1,393	2,537	2,592	3,478

La Tabla 4 muestra un ejemplo del uso del test de Quade sobre el marco experimental descrito.

Los rankings medios T_j pueden ser comparados con os rankings obtenidos por el test de Friedman clásico. En este caso, PDFC es el primero con ranking 1,393; NNEP y IS-CHC+1NN obtenien la segunda y tercera posición, con rankings 2,537 y 2,592, respectivamente; y el último es FH-GBML con ranking 3,487. El test de Quade comprueba si la suma de rankings ponderados S_j es significativamente diferente de 0:

$$A_2 = 24(24+1)(2\cdot 24+1)4(4+1)(4-1)/72 = 24500$$

$$B = \frac{1}{24}[(-332)^2+11^2+27{,}5^2+293{,}5^2] = 4068{,}479$$

$$T_3 = \frac{23 \cdot 4068{,}479}{24500 - 4068{,}479} = 21{,}967$$

Con 4 algoritmos y 24 conjuntos de datos, T_3 se distribuye de acuerdo a una distribución F con $4-1=3$ y $(4-1)\cdot(24-1)=69$ grados de libertad. El valor p calculado usando la distribución $F(3,69)$ es $4{,}28 \cdot 10^{-10}$, por lo que la hipótesis nula se rechaza con un alto nivel de significancia.

5. Procedimientos Posteriores de Identificación de Diferencias

Excepto el test de signos múltiple, el resto de procedimientos no identifican las diferencias existentes entre la propuesta y cada uno de los métodos comparados. Estos test se limitan a detectar la existencia o no de diferencias en todo el conjunto de resultados. Si las encuentra, habría que proceder con un test posterior de identificación de diferencias, que operan sobre la distribución normal. A continuación, se explican los cálculos necesarios para aproximar el valor z de una distribución normal a partir de la diferencias entre dos rankings para cada uno de los tests descritos anteriormente:

- Test de Friedman: En [3] podemos ver que la expresión para calcular z es

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

donde R_i, R_j son los rankings obtenidos por 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 [3]

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

donde $\hat{R}_i$, $\hat{R}_j$ son los rankings medios calculados por Friedman Alineado.

- Test de Quade: En [2], el estadístico para comparar dos algoritmos se proporciona usando la distribución *t de student*, pero podemos aproximarla a la distribución normal y calcular el valor z [7].

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

donde $T_i = \frac{W_i}{n(n+1)/2}$, $T_j = \frac{W_j}{n(n+1)/2}$, y W_i y W_j son los rankings sin ajuste de medias proporcionados por Quade. De hecho, T_i y T_j calculan las magnitudes medias correctas.

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 explicamos [5]. Una vez obtenido el valor z, es posible obtener el valor no ajustado p. Estos procedimientos, en definitiva, calculan el Valor Ajustado p (VAP) considerando la familia de hipótesis para cada pareja de algoritmos comparados, por lo que explicaremos cómo calcular el VAP para cada uno [8]. Usaremos la siguiente notación:

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

Los 4 procedimientos pueden definirse según su modo de calcular los VAPs de la siguiente manera:

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

La Tabla 5 muestra los resultados en la forma final de VAPs para el marco experimental considerado en este paper. Como podemos ver, este ejemplo es válido para observar ligeras diferencias de potencia entre los procedimientos estudiados. Si comparamos los VAPs obtenidos con el nivel de significancia que se establece a priori en todo estudio estadístico, podremos contabilizar el número de rechazos (si $VAP_i \leq \alpha$) y retenciones de hipótesis que se obtienen.

6. ¿Cómo Estimar la Diferencia de Rendimiento entre Dos Algoritmos?

Utilizando los datos resultantes de la ejecución de varios clasificadores sobre múltiples conjunto de datos, el investigador podría estar interesdo en la estimación de las diferencias entre el rendimiento de dos de ellos. Un procedimiento que busca este propósito es la **Estimación de Contraste** que supone que las diferencias esperadas entre rendimientos de algoritmos son las mismas a lo largo de los conjuntos de datos. Asumimos que cada

Tabla 5: Valores p ajustados (PDFC es el método control)

i	1	2	3
algoritmo	FH-GBML	IS-CHC + 1NN	NNEP
	Friedman		
p no ajustado	$5{,}69941 \cdot 10^{-5}$	0,05735	0,05735
p_{Bonf}	$1{,}70982 \cdot 10^{-4}$	0,17204	0,17204
p_{Holm}	$1{,}70982 \cdot 10^{-4}$	0,11469	0,11469
p_{Hoch}	$1{,}70982 \cdot 10^{-4}$	0,05735	0,05735
p_{Li}	$6{,}04577 \cdot 10^{-4}$	0,05735	0,05735
	Friedman Alineado		
p no ajustado	$2{,}32777 \cdot 10^{-7}$	0,02729	0,03032
p_{Bonf}	$6{,}98332 \cdot 10^{-7}$	0,08188	0,09097
p_{Holm}	$6{,}98332 \cdot 10^{-7}$	0,05459	0,05459
p_{Hoch}	$6{,}98332 \cdot 10^{-7}$	0,03032	0,03032
p_{Li}	$2{,}40057 \cdot 10^{-7}$	0,02738	0,03032
	Quade		
p no ajustado	$6{,}43747 \cdot 10^{-4}$	0,02163	0,02843
p_{Bonf}	$1{,}93124 \cdot 10^{-4}$	0,06490	0,08528
p_{Holm}	$1{,}93124 \cdot 10^{-4}$	0,04326	0,04326
p_{Hoch}	$1{,}93124 \cdot 10^{-4}$	0,02843	0,02843
p_{Li}	$6{,}62538 \cdot 10^{-4}$	0,02178	0,02843

medición de rendimiento se refleja como diferencias entre rendimientos de los algoritmos. Consecuentemente, estamos interesados en estimar el constraste entre medianas de muestras de resultados considerando todas las comparaciones pareadas. Obtiene una diferencia cuantitativa calculada mediante medianas entre dos algoritmos.

Procedemos 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 conjuntos de datos. En otras alabras, calculamos las diferencias

$$D_{i(uv)} = x_{iu} - x_{iv}$$

donde $i = 1, ..., n$; $u = 1, ..., k$; and $v = 1, ..., k$. Formamos parejas de rendimiento solo para aquellas parejas donde $u < v$.

2. Buscamos la mediana de cada conjunto de diferencias y la llamamos Z_{uv}. Llamamos Z_{uv} al *estimador no ajustado* de $M_u - M_v$. Puesto que $Z_{vu} = Z_{uv}$, tenemos que calcular solamente Z_{uv} en los casos donde $u < v$. Hay $k(k-1)/2$ medianas. Nótese que $Z_{uu} = 0$.

Tabla 6: Diferencias entre parejas de rendimientos en cada conjunto de datos para cada pareja de algoritmos

Conjunto	$D_{i(12)}$	$D_{i(13)}$	$D_{i(14)}$	$D_{i(23)}$	$D_{i(24)}$	$D_{i(34)}$
adult*	-0,021	-0,033	-0,043	-0,012	-0,022	-0,010
breast	-0,021	0,003	0,014	0,024	0,035	0,011
bupa	0,020	0,151	0,099	0,131	0,078	-0,053
car	0,133	0,114	0,203	-0,019	0,071	0,089
cleveland	-0,045	-0,067	-0,007	-0,021	0,039	0,060
contraceptive	-0,001	0,022	0,064	0,023	0,065	0,042
dermatology	0,096	0,014	0,436	-0,083	0,339	0,422
ecoli	0,024	0,012	0,063	-0,012	0,039	0,051
german	0,043	0,026	0,040	-0,017	-0,003	0,014
glass	0,137	0,040	0,101	-0,097	-0,036	0,062
haberman	-0,006	-0,003	-0,010	0,004	-0,003	-0,007
iris	0,020	0,013	0,007	-0,007	-0,013	-0,007
lymphography	0,080	0,031	0,141	-0,049	0,061	0,110
mushrooms*	0,006	0,515	0,087	0,509	0,081	-0,428
newthyroid	0,000	0,010	0,038	0,010	0,038	0,028
penbased*	0,029	0,049	0,352	0,020	0,323	0,302
ring*	0,205	0,145	0,130	-0,061	-0,076	-0,015
satimage*	0,067	0,012	0,075	-0,054	0,008	0,062
shuttle*	-0,019	-0,030	0,018	-0,011	0,038	0,048
spambase*	0,037	0,063	0,120	0,026	0,083	0,057
thyroid*	-0,013	-0,001	0,008	0,011	0,021	0,010
vehicle	0,194	0,235	0,282	0,041	0,089	0,047
wine	0,016	0,028	0,050	0,011	0,034	0,023
wisconsin	-0,001	-0,006	-0,006	-0,005	-0,005	0,000

3. Calculamos la media de cada conjunto de medianas no ajustadas que tienen el mismo prefijo y llamamos a dicho resultado m_u; esto es, calculamos

$$m_u = \frac{\sum_{j=1}^{k} Z_{uj}}{k}, u = 1, ..., k$$

4. El estimador de $M_u - M_v$ es $m_u - m_v$, donde u y v obtiene valores desde 1 a k. Por ejemplo, la diferencia entre M_1 y M_2 es $m_1 - m_2$.

El procedimiento se ilustra considerando el mismo marco experimental que en los estudios previos. La Tabla 6 muestra los cálculos realizados.

De los datos recogidos en la Tabla 6, encontramos que las seis medianas son $Z_{12} = 0{,}02$, $Z_{13} = 0{,}018$, $Z_{14} = 0{,}064$, $Z_{23} = -0{,}006$, $Z_{24} = 0{,}038$ y $Z_{34} = 0{,}035$. Ahora calculamos las medias para M_1 y M_2:

$$m_1 = \frac{0 + 0{,}02 + 0{,}018 + 0{,}064}{4} = 0{,}026$$

$$m_2 = \frac{-0{,}02 + 0 + (-0{,}006) + 0{,}038}{4} = 0{,}003$$

Nuestro estimador de $M_1 - M_2$ es $m_1 - m_2 = 0{,}026 - 0{,}003 = 0{,}023$. Es decir, la diferencia en acierto entre PDFC y NNEP estimada sobre multiples conjuntos de datos es igual a 0,023. La Tabla 7 muestra todos los estimadores entre los algoritmos.

Tabla 7: Estimación de contraste basada en medianas entre todos los algoritmos del estudio experimental

	PDFC	NNEP	IS-CHC+1NN	FH-GBML
PDFC	0,00000	0,02257	0,01976	0,05955
NNEP	-0,02257	0,00000	-0,00281	0,03698
IS-CHC-1NN	-0,01976	0,00281	0,00000	0,03979
FH-GBML	-0,05955	-0,03698	-0,03979	0,00000

7. Preguntas Frecuentes sobre el Uso de Tests no Paramétricos

En esta sección recogemos una serie de preguntas comunes acerca del uso de tests no paramétricos y planteamos sus respuestas:

¿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 en memoria.

¿Podemos comparar algoritmos determinísticos con estocásticos? Los tests no paramétricos pueden comparar ambos tipos de algoritmos 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 medios 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 provenientes de varias ejecuciones (al menos 3) 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 conjuntos de datos (dominios) debe ser superior al doble de algoritmos como mínimo. Con menos conjuntos, 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 conjuntos de datos es muy grande en relación al de algoritmos, los resultados tienden a ser ineficaces según el teorema central de límite [7]. Para un test de comparaciones pareadas, como el test de Wilcoxon, Demsar [4] recomendó un máximo de 30 conjuntos de datos. En comparaciones múltiples con método control, dependerá del número de algoritmos que se comparen, aunque podríamos indicar que utilizar un $n > 8 \cdot k$ 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 cuanto 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. Para ello existen los tests de múltiples comparaciones.

¿Podemos usar solo los valores de ranking obtenidos para justificar los resultados? Con los valores de rankings obtenidos con Friedman y derivados podemos establecer

una ordenación clara entre algoritmo e incluso medir las diferencias entre ellos, pero no se puede concluir que la 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 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 de Friedman? Las diferencias entre los tres métodos son pequeñas y muy dependientes de los resultados a analizar. Estudio teóricos demuestran que tanto el test de Friedman Alineado como el de Quade tienen mejor rendimiento cuando comparamos con pocos algoritmos, no más de 4. El test de Quade también supone un cierto riesgo porque asume que los data sets más relevantes son aquellos que presentan mayores diferencias entre los métodos, y esto no tiene por qué ser así.

¿Qué procemientos a posteriori deberían usarse? Consideramos que el test de Holm debe aparecer siempre en toda comparación, mientras que Bonferroni nunca por su conservatividad. El test de Hochberg y el de Li pueden servir de complemento cuando su uso permita rechazar más hipótesis de las que rechaza Holm. Toda hipótesis rechazada por cualquier procedimiento está correctaente 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 valores p no ajustados obtenidos en las hipótesis iniciales, y sólo cuando son menores que 0,5, el test obtiene su mejor rendimiento.

8. Comentarios Finales

En este trabajo presentamos todos los tests no paramétricos para comparar nuevas propuestas de Minería de Datos con ejemplos y recomendaciones sobre su uso. En el URL http://sci2s.ugr.es/sicidm se proporcionan más detalles y software para realizar los análisis estadísticos descritos en este trabajo.

Agradecimientos

Este trabajo ha sido financiado por TIN2008-06681-C06.

Referencias

[1] Alcalá-Fdez, J., Sánchez, L., García, S., del Jesus, M.J., Ventura, S., Garrell, J.M., Otero, J., Romero, C., Bacardit, J., Rivas, V.M., Fernández, J.C., Herrera, F. *KEEL: a software tool to assess evolutionary algorithms to data mining problems*, Soft Computing 13 (3) (2009) 307-318.

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

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

[4] Demšar, J. *Statistical comparisons of classifiers over multiple data sets*, Journal of Machine Learning Research 7 (2006) 1-30.

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

[6] García, S., Herrera, F. *An extension on "Statistical comparisons of classifiers over multiple data sets"*, Journal of Machine Learning Research 9 (2008) 2677-2694.

[7] Sheskin, D.J., *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, 2006.

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