Comparación estadística de algoritmos de aprendizaje estocásticos usando tests extendidos a datos intervalo-valorados y borrosos

José Otero
Departamento de Informática
Universidad de Oviedo
jotero@uniovi.es

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

Inés Couso
D. de Estadística e I. O. y D. M.
Universidad de Oviedo
couso@uniovi.es

Resumen

Cuando un diseño experimental basado en la validación cruzada se usa para comparar metaheurísticas o algoritmos evolutivos, es corriente repetir el aprendizaje varias veces por cada pareja de conjuntos entrenamiento – prueba. No obstante, como los resultados de las diferentes repeticiones no serán, en general, independientes, esta práctica puede producir conclusiones cuestionables.

En este trabajo se propone representar la información contenida en cada uno de los conjuntos de resultados asociados al mismo par entrenamiento – prueba mediante un intervalo, o mediante un conjunto borroso, y extender los tests estadísticos pertinentes a estos tipos de datos. De esta forma, la comparación por pares de estos conjuntos dará lugar un p-valor intervalo valorado o borroso, que contendrá información tanto acerca de las diferencias en promedio de los resultados del aprendizaje sobre los conjuntos de prueba como acerca de las diferencias entre las dispersiones esperadas en las ejecuciones de ambos algoritmos.

Palabras Clave: Diseño experimental, validación cruzada, algoritmos estocásticos, tests estadísticos para datos borrosos.

1. Introducción

La comparación de algoritmos de modelado y clasificación basada en la validación cruzada es empleada de forma habitual en Aprendizaje de Máquina, donde suele completarse con un test estadístico que juzga las diferencias entre la medias de dos muestras apareadas [7].

No obstante, si alguno de los algoritmos de aprendizaje se basa en una búsqueda randomizada, como ocurre en la mayoría de las metaheurísticas y en la práctica totalidad de los algoritmos evolutivos, cada ejecución del algoritmo produce distintos resultados aún para la misma combinación de conjuntos de entrenamiento y prueba. En esta situación, es práctica habitual repetir el aprendizaje varias veces por cada partición. Esto supone una modificación importante del diseño experimental, que potencialmente puede conducir a resultados cuestionables, dado que los resultados de las diferentes repeticiones no serán, en general, independientes. En el caso extremo, al comparar algoritmos deterministas con algoritmos estocásticos, no hay una posición clara en la literatura acerca de si se deben replicar los resultados de los algoritmos deterministas tantas veces como se ha hecho con los estocásticos (con lo que la muestra claramente no consta de elementos independientes) o bien se debe comparar cada uno de los resultados del algoritmo determinista con la media de todas las ejecuciones del algoritmo estocástico para la partición correspondiente, con lo que se pierde la información de la dispersión de los resultados de este último [9].

En la Figura 1 se ilustra este fenómeno. En horizontal se muestran los resultados por partición obtenidos mediante un algoritmo esto-

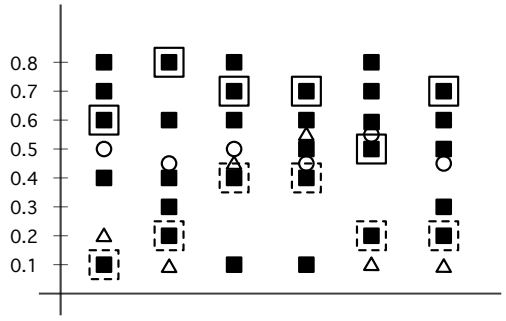

Figura 1: Ejemplo ilustrativo de diseño experimental VC seguido de promediado por partición. En vertical se representa el error, en horizontal los resultados por partición. Los cuadrados negros son los resultados obtenidos por un algoritmo estocástico, su media por partición se representa por círculos. Los triángulos son los resultados de un algoritmo determinista.

cástico (cuadrados negros). Estos valores se agregan en la media por partición (círculos). Por otra parte se tienen los resultados de un algoritmo determinista (triángulos). Si se efectúa un contraste de igualdad de medias usando como muestras los valores marcados por círculos y los marcados con triángulos, se rechazará la hipótesis nula. Sin embargo no se obtendrá ninguna información sobre la dispersión de los resultados: la media puede estar desviada debido a algún valor extremadamente bajo o alto, por ejemplo. Una alternativa sería usar todas las muestras posibles obtenidas tomando un valor del algoritmo estocástico de cada partición y obtener los p-valores máximo y mínimo. En el caso de la figura 1 se han marcado con cuadrados los valores del algoritmo estocástico que proporcionan esos valores. Sin embargo este análisis se reduce al "mejor" y "peor" caso. No se tiene ninguna información sobre cuanto son de representativos los valores de las muestras que proporcionan esos valores extremos dentro de los resultados obtenidos por partición por el algoritmo estocástico.

Lo más adecuado sería realizar un contraste en el que se utilizasen todos los resultados obtenidos para cada partición sin agregarlos en un único valor y de modo que se tenga en cuenta su distribución. Esto es especialmente importante por otra razón que no hemos tratado hasta el momento. En el caso más general, probablemente no se podrá utilizar un test paramétrico para realizar el contraste. La alternativa usual es entonces utilizar el test de Wilcoxon, por ejemplo. Este tipo de test, basado en rangos, no tiene en cuenta la distancia a la que se encuentran las muestras que se comparan; únicamente se tiene en cuenta cuándo uno de los valores de una de las muestras es mayor que el correspondiente en la otra muestra. Por ejemplo, si todos los valores de una muestra del error de un algoritmo son menores que los de otro algoritmo, el p-valor obtenido va a ser pequeño (se rechazará la hipótesis nula) sin importar *cuanto* menores sean.

1.1. Codificación borrosa de los conjuntos de resultados

Nuestro enfoque de este problema se enlaza con el tratamiento de los denominados "datos de baja calidad" [8]. Se entenderá que la estimación del error cometido por el algoritmo estocástico tiene asociada una tolerancia o, en otras palabras, que los resultados de las repeticiones de los algoritmos se corresponden con una hipotética serie de medidas que deberían haber arrojado idéntico valor. Cada una de estas series se representará con un conjunto borroso (o por un intervalo) lo que de acuerdo con diferentes autores [2, 6] es la codificación más adecuada para reconciliar resultados de varios experimentos que entran en conflicto.

En este estudio se empleará una codificación basada en la distribución bootstrap de la media de los datos [8], y se extenderán los tests de acuerdo con la metodología expuesta en [4]. En este caso, la extensión de un test estadístico a la comparación de muestras de variables aleatorias borrosas produce como resultado un p-valor borroso [3] que, como se verá, proporciona simultáneamente información acerca de las diferencias entre la calidad media de los algoritmos y acerca de cómo de dispersos están los resultados de estos (ver figura 2). Así pues, el objetivo último de este trabajo es la aplicación de ciertas extensiones borrosas de los tests estadísticos a la comparación de al-

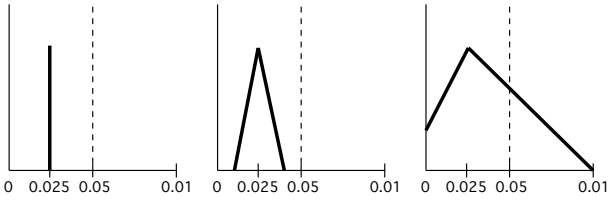

Figura 2: p-valores borrosos. Parte izquierda: las diferencias entre los errores medios de diferentes algoritmos son significativas si el p-valor es inferior a un valor acordado, p.e. 0.05. Centro y derecha: las extensiones borrosas de los tests producen conjuntos de p-valores que pueden estar completa o sólo parcialmente bajo el umbral. El punto modal coincide con la estimación crisp, pero la especificidad del p-valor indica en qué casos, como el mostrado más a la derecha, la dispersión de los resultados es demasiado elevada como para que la comparación sea decisiva.

goritmos estocásticos de aprendizaje con sus contrapartidas deterministas, con el fin de estudiar si existe alguna ganancia de información utilizando este enfoque. Usaremos para ello algunos datasets y algoritmos incluidos en el software KEEL [1].

En la siguiente sección se describe el problema de la aleatoriedad en la estimación del error de test en los algoritmos de modelado estocásticos y se citan las contribuciones más destacadas en el campo del análisis estadístico de datos imprecisos. En la sección 3 se detallan los experimentos realizados y se muestran los resultados obtenidos. Finalmente en la sección 4 recapitulamos sobre lo discutido en este trabajo.

2. Descripción del problema y solución propuesta

Supongamos que deseamos comparar el algoritmo estocástico E con el algoritmo determinista D. Para ello se puede aplicar validación cruzada (VC) con n subconjuntos. Como es sabido, se particiona el dataset S utilizado en n subconjuntos disjuntos $S_1, S_2, \ldots, S_n$, con $\bigcup_{i=1}^{n} S_i = S$. En la repetición i-ésima de la VC se usa $\bigcup_{j \neq i} S_j$ como conjunto de entrenamiento y S_i como conjunto de test. Como resultado, se obtienen dos muestras de tamaño n de las estimación del error de los algoritmo D y E. Llamaremos a estas muestras $\{d_1, ..., d_n\}$ y $\{e_1, ..., e_n\}$, respectivamente.

Supongamos que para el algoritmo E la VC se repite m veces desde puntos de partida elegidos al azar y manteniendo los subconjuntos de validación inalterados. En cada repetición $j = 1 \ldots m$ obtendremos una muestra $\{e_{1j}, ..., e_{nj}\}$ de la estimación del error del algoritmo. Generalmente se agregan estos resultados para obtener el error medio por partición,

$$e_{f_i} = \frac{1}{m} \sum_{j=1}^{m} e_{ij}, \tag{1}$$

y se comparan las muestras $\{d_i\}$ y $\{e_{f_i}\}$. Esta agregación conlleva una pérdida de información; por ejemplo, no sería posible distinguir entre un algoritmo con una tasa de error constante del 50 % y otro que tiene un error de 0 % en la mitad de los casos y del 100 % en el resto. En su lugar, nosotros proponemos construir n pertenencias borrosas $\widetilde{e_i}$, cada una de ellas compatible con los m valores $\{e_{ij}\}_{j=1,...,m}$.

2.1. Tests estadísticos extendidos

En lo relativo a la extensión de un test estadístico a datos borrosos, los trabajos más relavantes para nuestro enfoque son [4][5][6], donde se discuten diferentes extensiones de tests nítidos y, en menor medida, [3], donde se propone un método para defuzzificar p-valores borrosos. En cualquiera de estos casos, la naturaleza imprecisa de la estimación del error produce como resultado un p-valor impreciso. De igual modo, el test extendido a los datos imprecisos tendrá una potencia y un error de tipo I también imprecisos [10].

2.2. Solución propuesta

La extensión propuesta se basa en el método de Montecarlo, tal y como se sugiere en [6], y parte del test no paramétrico de Wilcoxon. Sus pasos son:

1. Escoger aleatoriamente un conjunto de n índices $\{i_1, i_2, \ldots, i_j, \ldots, i_n\}, i_j \in 1 \ldots m$, uno por cada una de las diferentes repeticiones del algoritmo estocástico en los n conjuntos de validación.

2. Calcular el p-valor pv correspondiente al test de Wilcoxon comparando las muestras del algoritmo estocástico E, $e_{i_1}, e_{i_2}, \ldots e_{i_n}$ con las del algoritmo determinista D, $d_1, d_2, \ldots, d_n$.

3. Inicializar los p-valores máximo y mínimo, pv_{min} y pv_{max}, con el valor obtenido.

4. Obtener una nueva muestra del error del algoritmo estocástico como en el paso 1, y calcular el p-valor pv correspondiente al test de Wilcoxon. Actualizar pv_{min} y pv_{max}

5. Si no se ha alcanzado el número de repeticiones especificado volver al paso 4.

Este procedimiento es fácilmente generalizable a distintos subconjuntos de las muestras del error por partición del algoritmo estocástico. Por ejemplo, se puede aplicar a los valores que estén comprendidos entre distintos cuantiles, de modo que se tiene entonces una percepción más completa de como incide la dispersión de los resultados en los p-valores máximo y mínimo. En la figura 3 se muestra un ejemplo similar al de la Figura 1, encerrando en un rectángulo los valores comprendidos dentro de los cuantiles 0.25 y 0.75. Como se puede observar, si se aplica el procedimiento anterior a los valores del algoritmo estocástico resaltados, nunca se va a encontrar un valor del algoritmo estocástico por debajo del valor correspondiente del algoritmo determinista, cosa que no sucedería si se considerasen todos los valores obtenidos para cada partición. El efecto que se consigue es restringir el contraste a los valores del

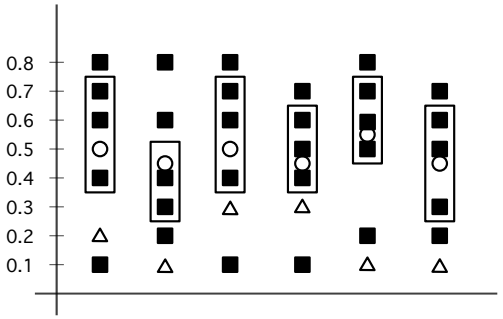

Figura 3: Ejemplo similar al de la Figura 1 resaltando las muestras del algoritmo estocástico comprendidas entre los cuantiles 0.25 y 0.75. Lógicamente el p-valor mínimo será mayor y el máximo menor que si se utilizasen todos los datos.

algoritmo estocástico más próximos a la mediana de la muestra. Se puede ver este procedimiento como intermedio entre considerar todos los valores o sólo la media. De esta forma se podría generalizar el proceso para una serie de valores $\{\alpha_1, \alpha_2, \ldots \alpha_i \ldots, \alpha_n\}$, con $\alpha_i < \alpha_j, \forall i < j$ y calcular los p-valores máximo y mínimo para los valores de la muestra del error del algoritmo estocástico entre los cuantiles $((1-\alpha_i)/2, 1-(1-\alpha_i)/2)$ para $i \in 1 \ldots n$. Se tiene entonces un conjunto de estimaciones de p-valores máximo y mínimo $\{(pv_{min,1}, pv_{max,1}), \ldots (pv_{min,i}, pv_{max,i}) \ldots \quad (pv_{min,n}, pv_{max,n})\}$ cumpliéndose que $pv_{min,i} > pv_{min,j}, \forall i < j$, $pv_{max,i} < pv_{max,j}, \forall i < j$, es decir, los intervalos de p-valores están anidados y por tanto tiene sentido emplear una función de pertenencia borrosa para representar de forma conjunta todos estos intervalos.

3. Estudio experimental

3.1. Metodología

Se comparará un algoritmo de regresión simbólica mediante recocido simulado [11] con la regresión lineal y cuadrática, usando varios datasets tomados de la herramienta KEEL [1]: "ele1""machine-CPU' "daily-elec" y "Friedman". Se usará validación cruzada con 10 sub-

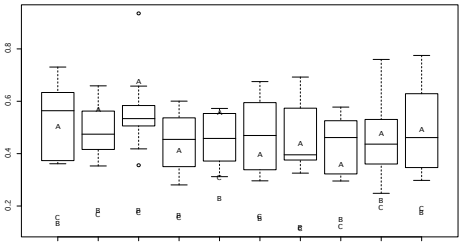

Figura 4: Boxplots de las dispersiones de los resultados del algoritmo SAP sobre el dataset "Daily". Se ha superpuesto la media de cada subconjunto para SAP (A), y los resultados por partición de "linear" (B) y "quadratic" (C).

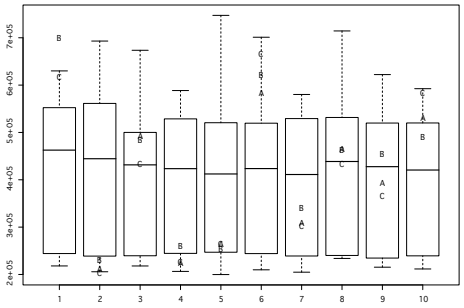

Figura 5: Boxplots de las dispersiones de los resultados del algoritmo SAP sobre el dataset "Ele1" en cada subconjunto de validación. Se ha superpuesto la media de cada subconjunto para SAP (A), y los resultados por partición de "linear" (B) y "quadratic" (C).

conjuntos, y el aprendizaje se repetirá 30 veces por cada subconjunto en los algoritmos randomizados.

Los resultados obtenidos se van a comparar de dos formas distintas. Por una parte, usando las muestras del error obtenidas se aplica el método explicado en la sección anterior y se calculan los p-valores máximo y mínimo mediante el test de Wilcoxon para las comparaciones de cada algoritmo estocástico con cada algoritmo determinista, usando todos los valores obtenidos en las 30 repeticiones y también los subconjuntos de muestras entre distintos cuantiles. El número de muestras utlizadas en el análisis Montecarlo es 10000. Por otra parte, se calcula la media del error de test de las 30 ejecuciones de validación cruzada para cada partición y se calculan los p-valores correspondientes a los mismos contrastes.

El objetivo final es observar si se obtienen las mismas conclusiones (rechazo o aceptación de la hipótesis nula del contraste) en los dos casos y si existe pérdida de información al agregar los datos de las particiones en la media.

3.2. Resultados

En primer lugar mostramos de forma gráfica los resultados, desglosados por subconjunto de validación, que han sido obtenidos por los distintos algoritmos sobre cada uno de los datasets utilizados. En la Figura 4 se muestran mediante boxplots los errores de las 30 repeticiones del algoritmo SAP sobre el dataset "Daily". Se han sobreimpreso en los boxplots las medias de cada subconjunto de validación de las 30 repeticiones de SAP junto con los resultados obtenidos con "Linear" y "Quadratic". En una primera inspección, parece existir una clara diferencia entre "Quadratic" y "Linear" frente a SAP.

En las figuras 5 y 6 se muestran de forma análoga los resultados sobre el dataset "Ele1" y "Friedman". En el primero de estos, al contrario que en el experimento anterior, los resultados nítidos están casi siempre comprendidos entre los cuantiles extremos de los boxplot. Asímismo, en el dataset "Friedman" hay diferencias relevantes entre uno de los algoritmos y el resto, en este caso "Quadratic". Finalmente, en la Figura 7, se da una situación en la que no se puede obtener una impresión clara de los resultados obtenidos por los distintos algoritmos sobre el dataset "Machine".

Los resultados que se han mostrado en las figuras anteriores se han utilizado para realizar un test como el propuesto en la sección 2, utilizando las muestras obtenidas en cada subconjunto de validación y un test clásico (el test de Wilcoxon) utilizando las medias por subconjunto de validación de SAP. Observe que un test extendido tiene tres conclusiones posibles; además de "rechazar" y de "no recha-

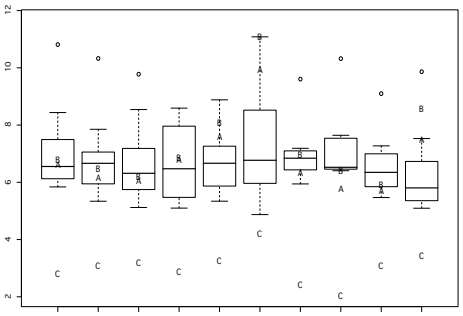

Figura 6: Boxplots de las dispersiones de los resultados del algoritmo SAP sobre el dataset "Friedman". Se ha superpuesto la media de cada subconjunto para SAP (A), y los resultados por partición de "linear" (B) y "quadratic" (C).

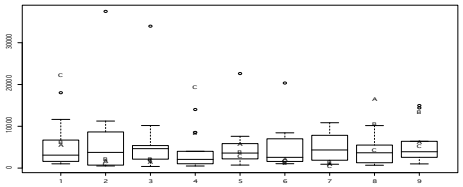

Figura 7: Boxplot por partición de los resultados del algoritmo SAP sobre el dataset "Machine". Superpuestos la media por partición para SAP (A), y los resultados por partición de "linear" (B) y "quadratic" (C).

zar" la hipótesis nula, existe la tercera posibilidad "datos no concluyentes", que en este estudio hemos asimilado a "no rechazar", ya que la muestra no contiene información como para afirmar que la hipótesis nula es improbable.

En la tabla 1 se muestran los p-valores máximos y mínimos junto con los p-valores nítidos obtenidos cuando se comparan los resultados de SAP y "Linear". Como se puede observar, los p-valores obtenidos son coherentes con la representación gráfica de los datos utilizados. Así en el caso del dataset "Daily" se rechaza la hipótesis nula con claridad y coinciden el p-valor máximo y mínimo. En el caso de "Ele1" también existe acuerdo entre los dos tests, puesto que el p-valor mínimo es mayor

	Daily	Ele1	Fried.	Mach.
min	1.0825e-05	0.4359	0.0039	0.0432
max	1.0825e-05	1	1	1
crisp	1.0825e-05	1	0.3150	0.9118

Cuadro 1: SAP vs. linear. Fila superior, p-valor mínimo. Segunda fila: p-valor máximo. Última fila: p-valor nítido.

incluso que 0.1. Sin embargo, en al caso de "Friedman" se observa que el test impreciso ha emitido el resultado "no concluyente", ya que el mínimo es menor que 0.01, es decir, en alguna de las repeticiones se han obtenido resultados claramente diferentes en el caso de SAP. En este caso, la utilización del test habitual habría producido resultados erróneos, ya que se habría concluído que no hay diferencias significativas entre SAP y el modelo lineal para el problema "Friedman", cuando la sitiuación real es que en algún caso SAP ha convergido a soluciones sustancialmente peores.

En el caso del dataset "Machine", existe consenso entre los dos tests para el nivel 0.1 pero no para 0.05 y 0.01. Para estos niveles el test no es concluyente, lo que de nuevo indica que limitarnos a la comparación entre las medias de las repeticiones no es suficiente, y el test crisp está produciendo una conclusión incorrecta que ha sido detectada por el test extendido. Finalmente, en la tabla 2 se muestran los p-valores correspondientes a la comparación entre el algoritmo SAP y "Quadratic". En el caso del dataset "Daily" existe consenso entre los dos tests a todos los niveles: se rechaza la hipótesis nula aún considerando el nivel 0.01 y el p-valor mínimo. Lo mismo sucede en el caso del dataset "Friedman". Para los otros dos datasets existe también consenso pero en el sentido contrario, en ningún caso se rechaza la hipótesis nula. Como conclusión, el uso del test extendido siempre ha servido para obtener conclusiones mejor fundadas, detectando las situaciones en que la media de los resultados de validación enmascara algoritmos que no han convergido a la solución correcta en todos los casos. En todos los casos se puede observar como el p-valor obtenido al contrastar las muestras del error del algoritmo determinis-

	Daily	Ele1	Fried.	Mach.
min	1e-05	0.3240	1e-05	0.1051
max	0.0001	1	4e-05	1
crisp	1e-05	0.9705	1e-05	0.4812

Cuadro 2: SAP vs. quadratic. Fila superior, p-valor mínimo. Segunda fila: p-valor máximo. Última fila: p-valor nítido.

ta con la muestra formada por las medias por subconjunto del algoritmo estocástico es intermedio entre los p-valores máximo y mínimo que se obtienen con los datos sin promediar.

Por último, en las tablas 3 y 4 se muestran los p-valores obtenidos siguiendo la segunda de las propuestas de la sección 2. Siguiendo la nomenclatura de dicha sección hemos elegido los valores 0.9, 0.75, 0.5, 0.25 para α en las expresiones que proporcionan los valores de los cuantiles entre los que estarán los valores de las muestras a considerar en el contraste.

En la tabla 3 se muestran los resultados correspondientes a la comparación de los algoritmos SAP y "linear". Como se puede observar, en el caso del dataset "Daily", los p-valores máximo y mínimo coinciden y por tanto no varían para los distintos cuantiles. Esto es un reflejo de la situación observada en los boxplot de la figura 4, en donde los valores del algoritmo "linear" (B) están siempre muy por debajo de la distribución de los valores obtenidos por el algoritmo SAP (boxplot y media A). Esta situación cambia para el dataset "Ele1" (representación gráfica en Figura 5), donde a la luz de los p-valores obtenidos no hay duda de que no se puede rechazar la hipótesis nula, igualdad de medias. En el caso del dataset "Friedman" (ver Figura 6), para el caso de los cuantiles más extremos, el p-valor mínimo cae por debajo del valor 0,05, lo cual quiere decir que en alguna de las repeticiones se han obtenido valores con SAP significativamente diferentes a los obtenidos con "linear" Los resultados obtenidos para "Machine", representados gráficamente en la Figura 7 son análogos a los obtenidos para "Ele1". Conclusiones similares se pueden extraer de la tabla 4, en donde se muestran los p-valores máximos y mínimos ob-

	Daily	Ele1	Fried.	Mach.
$pv_{0,9}^{min}$	1.082e-05	0.481	0.029	0.218
$pv_{0,75}^{min}$	1.082e-05	0.481	0.075	0.218
$pv_{0,5}^{min}$	1.082e-05	0.529	0.165	0.315
$pv_{0,25}^{min}$	1.082e-05	0.630	0.28	0.436
$pv_{0,25}^{max}$	1.082e-05	0.853	0.579	0.684
$pv_{0,5}^{max}$	1.082e-05	0.912	0.684	0.970
$pv_{0,75}^{max}$	1.082e-05	1.000	1.000	1.000
$pv_{0,9}^{max}$	1.082e-05	1.000	1.000	1.000

Cuadro 3: SAP vs. linear. En cada columna p-valores mínimos y máximos usando los valores comprendidos entre los cuantiles [0,05, 0.95], [0.125, 0.875], [0.25, 0.75], [0.375, 0.625], correspondientes a los valores de α 0.9, 0.75, 0.5, 0.25.

tenidos cuando se comparan por cuantiles los algoritmos SAP y "quadratic". En resumen, si el test extendido se basa en una representación borrosa por alfa cortes de los resultados de las repeticiones, a la información extra que proporciona el test cuando se emite el resultado "no concluyente" (esto es, que en algunas repeticiones el resultado no ha convergido a un valor compatible con el pretendido error medio) el menor nivel α al que aparece esta conclusión nos indica la fracción de casos en que el algoritmo tiene este comportamiento anómalo.

4. Conclusiones y trabajo futuro

En este trabajo se ha comprobado empíricamente que el uso de tests modernos, desarrollados originalmente para su uso con datos imprecisos, en combinación con una representación borrosa de los resultados de repetir un algoritmo de aprendizaje randomizado varias veces sobre el mismo conjunto de validación, permite obtener conclusiones más fuertes acerca de una experimentación. Se han visto ejemplos para los que las técnicas habituales habrían concluido erróneamente que una metaheurística no es sustancialmente peor que un algoritmo determinista, mientras que el test extendido es capaz de detectar falsas convergencias y, en su versión borrosa, de indicar cómo de frecuentes son estos casos, utilizando los mismos resulta-

	Daily	Ele1	Fried.	Mach.
$pv_{0,9}^{min}$	0.000	0.796	1.082e-05	0.315
$pv_{0,75}^{min}$	1.082e-05	0.796	1.082e-05	0.436
$pv_{0,5}^{min}$	1.082e-05	0.853	1.082e-05	0.481
$pv_{0,25}^{min}$	1.082e-05	0.912	1.082e-05	0.579
$pv_{0,25}^{max}$	1.082e-05	1.000	1.082e-05	0.912
$pv_{0,5}^{max}$	1.082e-05	1.000	1.082e-05	1.000
$pv_{0,75}^{max}$	1.082e-05	1.000	1.082e-05	1.000
$pv_{0,9}^{max}$	7.578e-05	1.000	1.082e-05	1.000

Cuadro 4: SAP vs. quadratic. En cada columna p-valores mínimos y máximos usando los valores comprendidos entre los cuantiles [0,05, 0.95], [0.125, 0.875], [0.25, 0.75], [0.375, 0.625], correspondientes a los valores de α 0.9, 0.75, 0.5, 0.25.

dos experimentales. A falta de un estudio más exhaustivo, el uso de este tipo de tests no ha producido resultados incoherentes en ningún caso y creemos que su adopción permitirá obtener conclusiones más estrictas sobre los datos sin apenas suponer un coste adicional en la computación, ya que el tiempo necesario para calcular los p-valores intervalo valorados o borrosos es, en general, muy inferior al usado en una sola de las evaluaciones del error del algoritmo.

Agradecimientos

Este trabajo ha sido financiado por el Ministerio de Ciencia e Innovación, proyecto TIN2008-06681-CO6-04

Referencias

[1] Alcalá-Fdez, et al. KEEL: A Software Tool to Assess Evolutionary Algorithms to Data Mining Problems. Soft Computing 13:3 (2009) 307-318.

[2] C. Bertoluzza, M. A. Gil, D. A. Ralescu (Eds.) Statistical Modeling, Analysis and Management of Fuzzy Data. Series: Studies in Fuzziness and Soft Computing , Vol. 87 2002, XIV

[3] I. Couso, L. Sánchez. Defuzzification of Fuzzy p-Values. *Advances in Soft Computing* (2008) 126-132

[4] I. Couso, L. Sánchez. Inner and outer fuzzy confidence regions determined by low quality data. Eurofuse Workshop 09: Preference Modelling and Decision Analysis

[5] T. Denoeux, M. H. Masson, P.A. Hubert, Nonparametric Rank-Based Statistics and Significance Test for Fuzzy Data, *Fuzzy Sets and Systems*, 153 (2005) 1-28

[6] S. Ferson, et al. Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty. Technical Report SAND2007-0939. Sandia National Laboratories, Albuquerque, New Mexico

[7] S. García, F. Herrera. An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets"for all Pairwise Comparisons. *Journal of Machine Learning Research* 9 (2008) 2677-2694

[8] L. Sánchez, I. Couso. Advocating the use of Imprecisely Observed Data in Genetic Fuzzy Systems. em IEEE Transactions on Fuzzy Systems 15:4 (2007) 551-562

[9] J. Otero, L. Sánchez. Diseños experimentales y tests estadísticos, tendencias actuales en Machine Learning. V Congreso Español sobre Metaheurísticas, Algoritmos Evolutivos y Bioinspirados (MAEB'07). Universidad de La Laguna. Puerto de La Cruz (España, 2007) 295-302

[10] J. Otero. Nuevos diseños experimentales para algoritmos de extracción de conocimiento con datos de baja calidad. XV Congreso Español sobre Tecnologías y Lógica Fuzzy. Punta Umbría, Huelva (España, 2010) 507-512.

[11] L. Sánchez, I. Couso, J.A. Corrales. Combining GP Operators with SA Search to Evolve Fuzzy Rule Based Classifiers. Information Sciences 136:1-4 (2001) 175-191.