Universidad&Ciencia
Pág. 64-82
Recibido: 30/11/2017 Aceptado: 19/02/2018
Publicado: 30/04/2018

Vol. 7, No. 2, abril-julio (2018)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

COMPARACIÓN DE CLASIFICADORES SOBRE MULTIPLES DATASETS CON PRUEBAS ESTADÍSTICAS NO PARAMÉTRICAS

COMPARISON OF CLASSIFIERS OVER MULTIPLE DATASETS WHITH NON-PARAMETRIC STATISTICAL TESTS

Autores: Orelvis López Jiménez[1]
Gladilis de la Caridad Barrera Garrido[1]
Carlos Rafael Bécquer Rodríguez[2]
Institución: [1] Universidad de Sancti Spíritus José Martí Pérez
[2] Grupo de Desarrollo. Joven Club de Computación de Sancti Spíritus
Correo electrónico: olopez@uniss.edu.cu

RESUMEN

En este trabajo se utiliza la herramienta Weka para medir el rendimiento de varios clasificadores, se compara el rendimiento de los mismos con varios datasets tomados de la Universidad de Ciencias Informáticas (UCI); demostrando que aplicando heurísticas para disminuir la dimensión de los datasets, así como la eliminación de ruidos en los mismos no afecta el rendimiento de los clasificadores. Este análisis se realiza aplicando pruebas estadísticas no paramétricas, el test de los rangos con signo de Wilcoxon para la comparación de dos o más 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. Alcanzando como resultado que las pruebas estadísticas no paramétricas son fiables para la comparación de los clasificadores y no afecta el rendimiento de los mismos una vez aplicadas las técnicas para la reducción de la complejidad de los datasets. Como resultado principal de esta investigación se puede generalizar este procedimiento para mejorar el rendimiento de clasificadores en otros datasets.

Palabras clave: Clasificadores, Comparación de clasificadores, Test no paramétricos, Aprendizaje automatizado.

Universidad&Ciencia
Pág. 64-82
Recibido: 30/11/2017 Aceptado: 19/02/2018
Publicado: 30/04/2018

Vol. 7, No. 2, abril-julio (2018)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

ABSTRACT

In this work Weka is used to measure the performance of several classifiers and compare their performance with several datasets taken from UCI and showing that applying heuristics to decrease the size of the datasets, as well as the elimination of noises in the same do not affect The performance of the classifiers. This analysis was performed using non-parametric statistical tests using the Wilcoxon signed rank test for comparison of two heuristics and, for the comparison of more than two heuristics, the Friedman test of multiple comparisons with the corresponding Namenyi and Bonferroni-Dunn a posteriori tests to establish the conclusions through the procedures of Holm and Hochberg. As a result, nonparametric statistical tests are reliable for comparison of classifiers and their performance is not affected once the techniques for reducing the complexity of the datasets are applied. With the results obtained at our discretion this procedure can be generalized to improve the performance of classifiers in other datasets

Keywords: Classifiers, Comparison of Classifiers, Non-Parametric Test, Machine Learning.

INTRODUCCIÓN

El aprendizaje puede ser definido como «cualquier proceso a través del cual un sistema mejora su eficiencia» (Felgaer et al., 2003). La habilidad de aprender es considerada como una característica central de los «sistemas inteligentes». En este trabajo se utiliza la herramienta de minería de datos Weka, acrónimo de Waikato Environment for Knowledge Analysis. La misma es una herramienta que permite el análisis y evaluación de las técnicas provenientes del aprendizaje automático. En los últimos años, el aprendizaje automatizado han aumentado las metaheurísticas y se 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(Garcia and Herrera, 2008).

En un trabajo típico de aprendizaje automatizado se aplican diversas heurísticas para el preprocesado de los datos y la eliminación de ruidos y errores de diversa índole asumiendo la hipótesis de que estas técnicas mejoran o mantienen el desempeño de los clasificadores. En otros trabajos experimentales se suele acudir al sentido común sobre la visualización de

Vol. 7, No. 2, abril-julio (2018)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

estos indicadores o con la aplicación de alguna técnica estadística se determina si la diferencia observada puede atribuirse al azar o son evidencia real de una diferencia que el rendimiento de los clasificadores. En otros estudios se justifica la aplicación de metherurísticas para disminuir la dimensión de los datasets y análisis de reducción de ruidos si estos no evidencian empeoramiento significativo en el indicador del rendimiento del clasificador. Todas estas conclusiones deben estar sustentadas en contrastes estadísticos rigurosos aplicados con imparcialidad en lugar de en la mera observación de tablas con indicadores de rendimiento.

En un estudio publicado por (Demsar, 2006) propone una guía para para el correcto análisis cuando se compara un conjunto de clasificadores sobre múltiples datasets. Demsar recomienda un conjunto de técnicas estadísticas no paramétricas (Zar, 1998, Sheskin, 2000) para la comparación de clasificadores bajo estas circunstancias. Además analiza el por qué estas técnicas son más convenientes que las técnica paramétricas. Otros estudias llevados a cabo por la guía de Demsar en el análisis del desempeño de clasificadores (C. Marrocco, 2008) plantea un nuevo procedimiento y lo compara con otros métodos por comparaciones de medias por pares.

En (Garcia and Herrera, 2008) plantea que el ranking calculado por el método de Friedman (Friedman, 1937) para la comparación de la significación de clasificadores. Demsar enfoca su trabajo en la introducción del test de Nemenyi para realizar comparaciones por pares (Nemenyi, 1936). García plantea que existen otros trabajos que se comparan el rendimiento de dos clasificadores con comparaciones de p-valor (Garcıa-Pedrajas and Fyfe, 2007). Test no paramétricos como Wilcoxon and Friedman calculan los p-valoers (Friedman, 1940, Wilcoxon, 1945). Sin embargo otros procedimientos como como es Holm (Holm, 1979), Hochberg (Hochberg, 1988), Hommel (Hommel, 1988) son usados para calcular los p-valores y la comparación de estos procedimientos no es muy difícil y (Garcia and Herrera, 2008) lo ponen de manifiesto.

MATERIALES Y MÉTODOS

Para llevar a cabo este estudio se selecciona cinco datasets tomadas del repositorio de aprendizaje automatizado UCI (UCI, 2017). Los nombres de las mismas son: Diabetes (Sigillito, 1990), Blocks Classification (Malerba, 1994), Vehicle Silhouettes (Mowforth and Shepherd, 1987), United States Congressional Voting Records Database (Schlimmer, 1984) y Wine (Blake, 1998) ver Tabla 1. A los cuales se aplican los clasificadores OneR, J48, IBK,

Vol. 7, No. 2, abril-julio (2018)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

SMO, MLP y Naive Bayes, tomados de la herramienta de aprendizaje automatizado Weka (Weka, 2012) en su versión 3.6. Se emplea también un framework experimental propuesto por (Garcia and Herrera, 2008) para el cálculo de las pruebas de hipótesis de los test estadísticos no paramétricos.

Dataset	Instancias	Atributos	Numéricos	Nominal	Class
Vote.arff	435	17	0	17	y,n
pima_diabetes.arff	768	9	8	1	0,1
page-blocks.arff	5473	11	10	1	1,2,3,4,5
vehicle.arff	946	19	18	1	opel,saab,bus,van
wine.arff	178	14	13	1	1,2,3

Tabla 1. Características de los datasets.

Clasificadores

Los clasificadores utilizados son una representación de diferentes tipos de clasificadores los cuales presentan buenos desempeños de manera general con diferentes tipos y estructuras de datasets.

Clasificador OneR

El clasificador OneR forma parte de los algoritmos basados en árboles de decisión, en los que el conocimiento obtenido en el proceso de aprendizaje se representa mediante un árbol. Dónde cada nodo interior contiene una pregunta sobre un atributo concreto y cada hoja del árbol se refiere a una de las posibles clasificaciones. Este algoritmo en concreto está basado en un árbol de profundidad 1, usando una única regla de decisión. Lo que hará será definir reglas simples a partir de las instancias. Así seleccionará el atributo con el menor error cuadrático medio para crear una regla a partir de un atributo, se elegirá la clase más frecuente de dicho atributo, la muestra que aparece más a menudo para un valor del atributo. En su implementación en Weka, este algoritmo elige la regla con el número más alto de instancias correctas, no la menor tasa de error. Presenta algunas inconvenientes como el «overfitting» de atributos nominales con valores únicos, la selección aleatoria de un atributo cuando las tasas de error son iguales o la selección aleatoria de una clase cuando varias clases dan la misma tasa de error con un atributo.

Universidad&Ciencia
Pág. 64-82
Recibido: 30/11/2017 Aceptado: 19/02/2018
Publicado: 30/04/2018

Vol. 7, No. 2, abril-julio (2018)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

Clasificador J48

Este clasificador en una implementación hecha por Weka basado en el algoritmo conocido como C4.5 (Quinlan, 1993), forma parte también de los algoritmos basados en árboles de decisión, al igual que el algoritmo anterior. La característica fundamental de este algoritmo es que incorpora una poda del árbol de clasificación una vez que éste ha sido inducido, una vez construido el árbol de decisión se podan aquellas ramas del árbol con menor capacidad predictiva (J. Han and Kamber, 2001). Este algoritmo es una mejora de ID3 (Quinlan, 1986), también basado en árboles, donde el criterio escogido para seleccionar la variable más informativa está basado en el concepto de cantidad de información mutua entre dicha variable y la variable clase.

Clasificador IBK

Este clasificador está basado en instancias, por ello consiste únicamente en almacenar los datos presentados. Cuando una nueva instancia es encontrada, un conjunto de instancias similares relacionadas es devuelto desde la memoria y usado para clasificar la instancia consultada. Se trata de un algoritmo del método *lazy learning*. Este método de aprendizaje se basa en que los módulos de clasificación mantienen en memoria una selección de ejemplos sin crear ningún tipo de abstracción en forma de reglas o de árboles de decisión. Cada vez que una nueva instancia es encontrada se calcula su relación con los ejemplos previamente guardados con el propósito de asignar un valor de la función objetivo para la nueva instancia. La idea básica sobre la que se fundamenta este algoritmo es que un nuevo caso se va a clasificar en la clase más frecuente a la que pertenecen sus K vecinos más cercanos. De ahí que sea también conocido como método K-NN: *K* Nearest Neighbours *(*Mitchell*, 1997).*

Clasificador SMO

Este clasificador está basado en redes neuronales, cuya característica más importante es su capacidad de aprender a partir de ejemplos, lo cual les permite generalizar sin tener que formalizar el conocimiento adquirido. El mismo se caracterizará por tener un aprendizaje no supervisado competitivo y por no tener ningún resultado objetivo al que la red deba tender. Además, SMO divide el problema en una serie de problemas más pequeños que se resuelven de forma analítica. Cada neurona de la red calcula la similitud entre el vector de entrada y su propio vector de pesos según un criterio de similitud establecido. A

Vol. 7, No. 2, abril-julio (2018)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

continuación, simulando un proceso competitivo, se declara vencedora la neurona cuyo vector de pesos es el más similar al de entrada. Esto hace que la red SMO se comporte como un clasificador, la neurona de salida activada representará la clase a la que pertenece la información de entrada.

Clasificador MLP

El clasificador MLP es una red de alimentación hacia adelante que consta de una capa de entrada, una capa de salida y una o más capas ocultas. El número de unidades de procesamiento en la capa de entrada es igual a la dimensión del vector de entrada. El número de unidades de procesamiento en la capa de salida es el mismo que el número de las clases destinados al sistema de clasificación. Las unidades de tratamiento entre dos capas consecutivas están totalmente conectadas con enlaces ponderados. Así como la regresión lineal se utiliza para encontrar el conjunto de coeficientes que produzcan el mínimo error, retropropagación con descenso de gradiente se propone para reducir al mínimo el error cuadrático medio entre la salida real de la MLP y la salida deseada. Una descripción con-precisa de que el algoritmo se puede encontrar en (Lippmann, 1987) o en (Richards, 1993). Una de las mayores dificultades de aplicar MLP es, como determinar la estructura optima de la red (cuantas capas ocultas y cuantas neuronas). En (Lippmann, 1987)se sugirió que un MLP sin más de dos las capas ocultas deben ser bastante para generar un límite de decisión arbitrariamente complejo. Con respecto al número óptimo de nodos ocultos para cada una de las capas ocultas, algunas sugerencias se dieron por (Lippmann, 1987) y (Pao, 1989). Sin embargo, la pregunta no tiene una respuesta satisfactoria todavía. Le basta mencionar aquí que no hay todavía ninguna teoría fuerte para tratar con este problema, y las estructuras correctas de MLPS tienen que ser encontradas por los experimentos.

Clasificador Naive Bayes

Este Los clasificadores Bayesianos (Duda and Hart, 1973) son clasificadores estadísticos, que pueden predecir tanto las probabilidades del número de miembros de clase, como la probabilidad de que una muestra dada pertenezca a una clase particular. La clasificación Bayesiana se basa en el teorema de Bayes y los clasificadores Bayesianos han demostrado una alta exactitud y velocidad cuando se han aplicado a grandes bases de datos Diferentes estudios comparando los algoritmos de clasificación han determinado que un clasificador

Universidad&Ciencia
Pág. 64-82
Recibido: 30/11/2017 Aceptado: 19/02/2018
Publicado: 30/04/2018

Vol. 7, No. 2, abril-julio (2018)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

Bayesiano sencillo conocido como el clasificador «naive Bayesiano» (John, 1997) es comparable en rendimiento a un árbol de decisión y a clasificadores de redes de neuronas. El clasificador naive Bayesiano se utiliza cuando se quiere clasificar un ejemplo descrito por un conjunto de atributos en un conjunto finito de clases. Clasificar un nuevo ejemplo de acuerdo con el valor más probable dados los valores de sus atributos. Además, el clasificador naive Bayesiano asume que los valores de los atributos son condicionalmente independientes dado el valor de la clase. Los clasificadores naive Bayesianos asumen que el efecto de un valor del atributo en una clase dada es independiente de los valores de los otros atributos. Esta suposición se llama «independencia condicional de clase». Ésta simplifica los cálculos involucrados y es considerado «ingenuo». Esta asunción es una simplificación de la realidad. A pesar del nombre del clasificador y de la simplificación realizada, el naive Bayesiano funciona muy bien, sobre todo cuando se filtra el conjunto de atributos seleccionado para eliminar redundancia, con lo que se elimina también dependencia entre datos.

Prueba de rangos con signo de Wilcoxon

La prueba de los rangos con signo de Wilcoxon (Wilcoxon, 1945) es una prueba no paramétrica para comparar la mediana de dos muestras relacionadas y determinar si existen diferencias entre ellas. Se utiliza como alternativa a la prueba T de Student cuando no se puede suponer la normalidad de dichas muestras. 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 clasificador. Formalmente, sea rang(d_i) la posición que ocupa | d_i | en una ordenación llamada rango de d_i. Sea R^+ y R^- las sumas respectivas de los rangos de las diferencias positivas y negativa. Los rangos d las diferencias nulas $d_i = 0$ se reparten entre ambas sumas y si existe un número impar de ellas se ignora una. Por tanto ver ecuaciones 1 y 2:

$$R^+ = \sum_{di>0} rang(d_i) + \frac{1}{2} * \sum_{di=0} rang(d_i) \quad (1) \qquad R^-$$

$$= \sum_{di<0} rang(d_i) + \frac{1}{2} * \sum_{di=0} rang(d_i) \quad (2)$$

Sea $T = min\,(R^+, R^-)$ si la hipótesis nula es cierta T debe aproximarse a $n(n+1)/4$ y cuanto más diferente sea el rendimiento de los dos clasificadores menor será T. Muchos

Universidad&Ciencia
Pág. 64-82
Recibido: 30/11/2017 Aceptado: 19/02/2018
Publicado: 30/04/2018

Vol. 7, No. 2, abril-julio (2018)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

textos de estadística incluyen valores críticos exactos para T para n hasta 25. Para un número mayor de instancias, el estadístico tiene distribución aproximadamente normal. El test de los rangos con signo de Wilcoxon es más sensible que la prueba T de Student, requiere conmensurabilidad de las diferencias, pero solo 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 ello disminuirá la potencia del test debido al aumento 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 la prueba T de Student. ndencia entre datos.

Prueba Friedman

La prueba de Friedman (Friedman, 1937) es una prueba no paramétrica desarrollado por el economista Milton Friedman. Equivalente a la prueba ANOVA (Fisher, 1959) para medidas repetidas en la versión no paramétrica, el método consiste en ordenar los datos por filas o bloques, reemplazándolos por su respectivo orden. Al ordenarlos, debemos considerar la existencia de datos idénticos. 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_i^j el rango del j-ésimo algoritmo sobre la i-ésima instancia. El test de Friedman compara los rangos medios de los algoritmos calculados por la ecuación 3. Bajo la hipótesis nula H_0 del rendimiento de todos los algoritmos es el mismo y los rangos R_j deben ser similares. El estadístico propuesto por Friedman en la ecuación 4 se distribuye de acuerdo a una X_F^2 con k-1 grados de libertad, cuando *n* y *k* son suficientemente grandes, como regla experimental *n* > 10 y *k* >5. Para un número menor de caso y algoritmos se han calculado los valores críticos exactos (Zar, 1998, Sheskin, 2000).

$$R_j = \frac{1}{n}\sum_{i=1}^{n} r_i^j \quad (3) \qquad X_F^2 = \frac{12n}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right] \quad (4) \qquad F_F = \frac{(n-1)X_F^2}{n(k-1) - X_F^2} \quad (5)$$

Universidad&Ciencia
Pág. 64-82
Recibido: 30/11/2017 Aceptado: 19/02/2018
Publicado: 30/04/2018

Vol. 7, No. 2, abril-julio (2018)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

Iman y Davenport mostraron en (Iman and Davenport, 1980) un mejor estadístico para X_F^2 que se distribuye de acuerdo con una distribución F de Fisher-Snedecor con $k-1\ y\ (k-1)(n-1)$ grados de libertad el F_F véase en la ecuación 5.

Pruebas 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 (Nemenyi, 1936) se usa para comparar todos los algoritmos entre sí. El rendimiento de dos algoritmos es significativamente diferente si los rangos promedios difieren, al menos en el valor crítico $CD = q_\alpha \sqrt{k(k+1)/6n}$ donde los valores críticos de q_α son los del estadístico de rangos estandarizados 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. El estadístico para comparar los rendimientos de los clasificadores i-ésimos y j-ésimos es Z ver ecuación 6. 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 α para compensar las comparaciones múltiples.

$$Z = \frac{R_i - R_j}{\sqrt{\frac{k(K+1)}{6n}}} \qquad (6)$$

El test de Bonferroni-Dunn (Dunn, 1961) divide el valor de ® 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 α = (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. 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 p1, p2,… pk de forma tal que p1 < p2 <… pk-1. Dos métodos simples usuales son el método descendente de Holm (Holm, 1979) y el ascendente de Hochberg (Hochberg, 1988). Ambos comparan cada p_i con $\alpha/(k-i)$, pero difieren en el orden de los tests. Según Demsar 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

Vol. 7, No. 2, abril-julio (2018)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

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 en (Demsar, 2006) se describen los procedimientos de Holm y Hochberg como métodos para realizar comparaciones a posteriori para el test 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.

RESULTADOS Y DISCUSIÓN

En todos los casos se realiza 10-fold 10 Cros Validation, para un total de 100 corridas por cada algoritmo. Primeramente se efectúa una corrida de todos los datasets sin realizar modificaciones contra todos los clasificadores para verificar el rendimiento de los mismos. En todo los casos se trabaja como medida de evaluación de los clasificadores con el porciento de bien clasificados. Se obtuvo los siguientes resultados ver Tabla 2. Evidenciando que todos los clasificadores alcanzaron resultados aceptables, como promedio fue superior al 78 %.

Dataset\Clasificador	rules.OneR	trees.J48	lazy.IBk	SMO	MLP	NaiveBayes
Vote.arff	95,63	96,57	93,08	95,77	94,43	90,02
pima_diabetes.arff	72,78	73,82	72,65	77,34	75,39	76,3
page-blocks.arff	93,78	96,87	95,92	77,34	96,23	90,84
vehicle.arff	51,41	72,45	71,51	74,34	81,67	44,79
wine.arff	77,52	93,82	94,94	98,31	97,19	96,62
Promedio	78,224	86,706	85,62	84,62	88,982	79,714

Tabla 2. Resultado corrida de clasificadores con dateset sin modificación.

Optimización de los datsets y selección de atributos.

El pre procesamiento de los datos se realiza con el fin de disminuir la complejidad del dataset con el fin de disminuir el costo computacional de los clasificadores empleados para el aprendizaje siempre que el resultado de ello no afecte la clasificación. Para ello se aplican pruebas estadísticas no paramétricas descritas en la sección 2.1.7 al 2.1.9 para demostrar que la optimización del dataset no afecta significativamente el resultado de los clasificadores. Ninguno de los dataset poseía valores ausentes por lo que no se le aplicaron

Vol. 7, No. 2, abril-julio (2018)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

técnicas para sustitución de valores ausentes. Luego se procede a eliminar los atributos con menos relevancia a la hora de clasificación para disminuir el costo computacional del aprendizaje. Weka posee filtros para la selección de atributos. Para medir la relevancia de los atributos se aplica filtros con 10-fold cros validation para minimizar errores. A todos se le aplican los filtros de selección de atributos que aparecen en la Tabla 3. Cada filtro con los valores por defecto que aparecen en Weka.

Identificador	Filtro heuristico	Dirección de búsqueda
A	CfsSubsetEval	BestFirst -D 1 -N 5
B	ChiSquaredAttributeEval	Ranker -T -1.7976931348623157E308 -N -1
C	GainRatioAttributeEval	Ranker -T -1.7976931348623157E308 -N -1
D	InfoGainAttributeEval	Ranker -T -1.7976931348623157E308 -N -1
E	ReliefFAttributeEval	Ranker -T -1.7976931348623157E308 -N -1
F	SVMAttributeEval	Ranker -T -1.7976931348623157E308 -N -1

Tabla 3. Filtros de selección de atributos con su dirección de búsqueda.

Después de aplicarle los filtros a las bases de datos se obtuvo como resultado lo que se muestra en la Tabla 5, la cual contiene en sus columnas el número del atributo y el orden que cada filtro le otorgó. De estos resultados se seleccionó los atributos que engloban más del 80% de relevancia para cada clasificador y por el criterio de selección de cuales atributos repetían en los primeros valores del ranking se realiza la selección de los que aparecen en la Tabla 4. El último en cada caso corresponde a la clase.

Dataset	Atributos seleccionados
Vote.arff	4,3,5,15,9,12,14,8,17
pima_diabetes.arff	1,2,6,7,8,9
page-blocks.arff	1,4,5,6,7,10,11
vehicle.arff	1,2,3,5,7,8,9,10,11,12,17,18,19

Universidad&Ciencia
Pág. 64-82
Recibido: 30/11/2017 Aceptado: 19/02/2018
Publicado: 30/04/2018

Vol. 7, No. 2, abril-julio (2018)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

wine.arff	1,4,6,7,10,11,12,13,14

Tabla 4. Listado de atributos seleccionados para cada dataset.

Luego que nos quedamos con los atributos más significativos realizamos un análisis para determinar en cada caso cuales instancias tenían valores a nuestro criterio que podría introducir ruido en la clasificación, para ello nos apoyamos en el análisis de histogramas para cada atributo y revisamos las medidas centrales como la media y la desviación estándar ya que la desviación estándar de un conjunto de datos es una medida de cuánto se desvían los datos de su media. El criterio que seguimos fue que los valores que estuviera más alejados que el valor de la media más o menos 3 veces la desviación estándar como un criterio para eliminar el ruido en los datos. Después de realizar estos análisis se volvió a aplicar los mismos clasificadores primero con los dataset con menos atributos y otra corrida con los dataset con menos atributos y con la limpieza de outlayer.

		Ranking de los atributos																	
Dataset	Heuristica	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
diabetes	A	2	6	8	7	5	1	3	4										
	B	2	8	6	5	1	7	4	3										
	C	2	6	8	1	5	7	4	3										
	D	2	6	8	5	1	4	7	3										
	E	2	6	4	1	8	7	3	5										
	F	2	6	1	7	3	8	5	4										
page-blocks	A	1	4	5	6	7	10	2	3	8	9								
	B	4	1	2	5	10	9	3	8	7	6								
	C	1	4	7	10	5	6	2	3	9	8								
	D	1	4	5	7	10	6	3	2	9	8								
	E	6	5	2	4	10	1	9	8	3	7								
	F	1	4	8	3	10	2	7	9	6	5								
Vehicle	A	6	7	11	12	14	15	16	5	8	3	4	9	1	2	10	13	17	18
	B	12	7	8	9	11	3	6	2	1	10	4	13	14	17	18	4	16	15
	C	11	12	7	9	8	6	4	3	13	5	18	14	2	1	16	10	15	17
	D	12	7	8	11	9	3	6	4	2	1	13	10	14	17	18	5	16	15
	E	8	18	7	12	9	3	10	11	2	17	1	13	4	6	14	15	5	16

Universidad&Ciencia
Pág. 64-82
Recibido: 30/11/2017 Aceptado: 19/02/2018
Publicado: 30/04/2018

Vol. 7, No. 2, abril-julio (2018)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

	F	3	8	4	14	17	5	10	18	1	7	13	12	2	11	6	9	15	16
Wine	A	1	2	3	4	5	7	10	11	12	13	6	8	9					
	B	7	13	10	12	11	1	6	2	9	4	5	8	3					
	C	12	7	10	13	1	6	11	3	4	9	5	2	8					
	D	7	13	10	12	1	11	6	2	5	9	4	8	3					
	E	12	7	13	1	10	6	11	2	8	9	4	5	3					
	F	13	10	7	12	1	11	4	3	2	8	9	6	5					
Vote	A	4	3	10	11	12	9	14	13	15	5	6	7	16	10	2	1		
	B	4	3	5	12	8	9	14	13	15	7	6	1	11	16	10	2		
	C	4	3	5	12	14	8	9	13	15	7	6	1	11	16	10	2		
	D	4	3	5	12	14	8	9	13	15	7	6	1	11	16	10	2		
	E	4	14	11	3	5	15	10	12	2	1	13	6	9	8	7	16		
	F	4	11	3	9	10	15	8	7	2	5	12	16	6	13	14	1		

Tabla 5. Orden de los atributos con los filtros de selección.

Luego de verificar los resultados obtenidos ver Tabla 6, se determinó que este procedimiento al no ser en el dataset Vehicle, la última corrida de los clasificadores fue mejor que la primera, demostrando que se pueden alcanzar iguales o mejores resultados con menos atributos que los iniciales que traían los datasets. En el caso del dataset vehicle se aprecia una mejora significativa de la segunda corrida con respecto a la primera y la última, pero no así entre la primera y la última puesto que son valores similares en todos los clasificadores.

Dataset	rules.OneR	trees.J48	lazy.IBk	SMO	MLP	NaiveBayes
Vote.arff	95.63	96.57	93.08	95.77	94.43	90.02
Vote-Atr.arff	96.06	96.06	94.95	96.06	94.74	91.42
Vote-Atr-Outl.arff	96.06	96.91	94.55	96.4	95.29	90.79
pima_diabetes.arff	72.78	73.82	72.65	77.34	75.39	76.3
pima_diabetes-Atr.arff	72.78	74.21	74.21	77.08	75.01	76.69
pima_diabetes-Atr-Out.arff	72.78	74.24	73.18	76.61	74.1	76.22
page-blocks.arff	93.78	96.87	95.92	77.34	96.23	90.84

Vol. 7, No. 2, abril-julio (2018)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

page-blocks-Atr.arff	93.78	97.11	95.74	92.39	96.01	94.59
page-blocks-Atr-Outl.arff	94.62	97.31	96.76	95.99	96.84	94.26
vehicle.arff	51.41	72.45	71.51	74.34	81.67	44.79
vehicle-Atr.arff	79.19	72.22	72.57	72.69	79.19	79.19
vehicle-Atr-Out.arff	53.16	69.77	71.56	71.32	78.97	44.56
wine.arff	77.52	93.82	94.94	98.31	97.19	96.62
wine-Atr.arff	77.52	94.38	96.62	96.06	97.19	96.06
wine-Atr-Out.arff	77.52	94.38	96.62	96.06	97.19	96.06

Tabla 6. Resultado de los clasificadores en las 3 corridas.

Comparación de los clasificadores.

Los clasificadores se utilizaron con 5 datasets y a cada uno se le realizaron 2 modificaciones teniendo como resultado que cada clasificador se corrió contra 15 dataset. Como resultado final se determina cuáles fueron los mejores clasificadores. 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. En este caso X^2_F = 12.99 con 5 grados de libertad y F_F = 2. Con 6 algoritmos y 15 datasets, F_F se distribuye de acuerdo a la distribución F con 5 y 70 grados de libertad. El valor crítico de F (3; 39) al nivel α = 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 = 0.686. Si no se distingue individualmente ningún algoritmo, se usa el test de Nemenyi para las comparaciones por pares. El valor crítico es α = 2.85 y la correspondiente diferencia crítica es CD = 2.85 x SE = 1.95. 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 α = 0.10 la diferencia crítica es CD = 2.589 x SE = 1.768. Se puede identificar 3 grupos de algoritmos que el rendimiento es similar MLP y SMO, J48 e IBK y el tercer grupo NaiveBayes y Rules OneR, ver Tabla 7. Estadísticamente no se concluye que se tienen pruebas para demostrar que MLP es superior a SMO. Pero si son superiores al resto para estos conjuntos de datos.

Vol. 7, No. 2, abril-julio (2018)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

Algoritmo	Ranking
MLP	2.6666666666666674
SMO	2.8000000000000003
trees.J48	3.0333333333333337
lazy.IBk	3.766666666666667
NaiveBayes	4.2
rules.OneR	4.533333333333334

Tabla 7. Ranking promedio de los algoritmos por Friedman.

Las mejoras realizadas a los datasets empíricamente están correctas pero cuanto de ello es cierto estadísticamente para ello se ajustan los datos como aparecen en la Tabla 8 y se aplica la prueba de Friedman. Según los resultados del test de Friedman con 2 grados de libertad y la hipótesis nula que las distribuciones de las tres modificaciones son las mismas se acepta la hipótesis nula y se demuestra que son iguales. Realizando análisis a posteriori se aplica el test de Nemenyi para la comparación por pares el valor crítico para $q_\alpha = 2.343$ para 3 instancias y su SE = 0.285 y su correspondiente diferencia critica es CD = 2.343 x SE = 0.604. Ya que la diferencia entre el mejor y el peor rendimiento es ya menor que el valor de CD se concluye que este test no es suficientemente potente para detectar diferencias entre los datos o que no existen datos suficientes. Por lo que se plantea otra hipótesis, es posible mejorar o mantener el rendimiento de los algoritmos si se realizan mejoras en los datos. La forma más fácil de verificar esto es calcular la diferencia critica con el test de Bonferoni-Dunn. En la tabla correspondiente se encuentra que el valor critico q_α para $\alpha = 0.05$ y 3 variantes es de 2.24 y la DC = 2.24 x SE = 0.56 y la diferencia del mejor y el pero es inferior a este valor por lo que se concluye que las variantes no son significativamente diferentes en cuanto al rendimiento de los algoritmos.

Algoritmo-dataset	Dataset S/M	Dataset M-1	Dataset M-2
oneR-V	95.63	96.06	96.06
oneR-d	72.78	72.78	72.78
oneR-p	93.78	93.78	94.62
oneR-ve	51.41	79.19	53.16
oneR-w	77.52	77.52	77.52

Vol. 7, No. 2, abril-julio (2018)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

NaivBayes-V	90.02	91.42	90.79
NaivBayes-d	76.3	76.69	76.22
NaivBayes-p	90.84	94.59	94.26
NaivBayes-ve	44.79	79.19	44.56
NaivBayes-w	96.62	96.06	96.06
IBK	93.08	94.95	94.55
IBK	72.65	74.21	73.18
IBK	95.92	95.74	96.76
IBK	71.51	72.57	71.56
IBK	94.94	96.62	96.62
MLP	94.43	94.74	95.29
MLP	75.39	75.01	74.1
MLP	96.23	96.01	96.84
MLP	81.67	79.19	78.97
MLP	97.19	97.19	97.19
J48	96.57	96.06	96.91
J48	73.82	74.21	74.24
J48	96.87	97.11	97.31
J48	72.45	72.22	69.77
J48	93.82	94.38	94.38
SMO	95.77	96.06	96.4
SMO	77.34	77.08	76.61
SMO	77.34	92.39	95.99
SMO	74.34	72.69	71.32
SMO	98.31	96.06	96.06

Tabla 8. Resultado calificador por dataset y sus modificaciones.

CONCLUSIONES

Cada vez son menos las dificultades para aplicar técnicas estadísticas a la hora de comparar el rendimiento de clasificadores y la aplicación de técnicas no paramétricas frente a las paramétricas se eliminan los supuestos necesarios a la hora de aplicar estas últimas. Además se concluye que el test de Friedman es una herramienta eficiente para seleccionar de manera certera cual es el mejor clasificador entre varios y que las pruebas a posteriori

Vol. 7, No. 2, abril-julio (2018)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

sirven de apoyo para mejorar el alcance de estas pruebas. También que la limpieza de los datasets en la manera que se realizó en el trabajo no influye negativamente en el resultado de los clasificadores y si en mejora en cuanto a tiempo de cómputo y limpieza de atributos.

BIBLIOGRAFÍA CONSULTADA

BLAKE, C.: Wine recognition. *In:* Forina, M. E. A., Parvus - *An extendible package for data exploration, classification and correlation. institute of pharmaceutical and food analysis and technologies, via brigata salerno*, 16147 Genoa, Italy, 1998.

MARROCCO, C. Y TORTORELLA, F.: «Maximizing the area under the ROC curve by pairwise feature combination», *Pattern Recognition,* Vol. 41**,** pp. 1961–1974, 2008.

DEMSAR, J.: «Statistical comparisons of classifiers over multiple data sets», *Journal of Machine Learning Research,* Vol.7**,** pp.1-30, 2006.

DUDA, R. Y HART, P.: *Pattern Classification and Scene Analysis,* NY, John Wiley & Sons, 1973.

DUNN, O.J.: «Multiple comparisons among means», *Journal of the American Statistical Association,* Vol.56**,** pp.52–64, 1961.

FELGAER, P., BRITOS, P.; SICRE, J.; SERVETTO, A.; GARCÍA-MARTÍNEZ, R. Y PERICHINSKY, G.: *Optimización de Redes Bayesianas Basada en Técnicas de Aprendizaje por Instrucción. VIII Congreso Argentino de Ciencias de la Computación*, 2003.

FISHER, R.A.: *Statistical methods and scientific inference (2nd edition),* New York, Hafner Publishing Co, 1959.

FRIEDMAN, M.: «The use of ranks to avoid the assumption of normality implicit in the analysis of variance», *Journal of the American Statistical Association,* Vol.32**,** pp.675–701, 1937.

FRIEDMAN, M.A.: «Comparison of alternative tests of significance for the problem of m rankings», *Annals of Mathematical Statistics,* Vol.11**,** pp.86–92, 1940.

GARCIA PEDRAJAS, N. Y FYFE, C.: «Immune network based ensembles», *Neurocomputing,* Vol.70, Num.7-9**,** pp.1155–1166, 2007.

GARCIA, S. Y HERRERA, F.: «An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons», *Journal of Machine Learning Research,* Vol.9, 2008.

HOCHBERG, Y.: «A sharper bonferroni procedure for multiple tests of significance». *Biometrika,* Vol.75**,** pp.800–802, 1988.

Vol. 7, No. 2, abril-julio (2018)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

HOLM, S.: «A simple sequentially rejective multiple test procedure», *Scandinavian Journal of Statistics,* Vol.6, pp.65–70, 1979.

HOMMEL, G.: «A stagewise rejective multiple test procedure», B*iometrika,* Vol.75, pp.383–386, 1988.

IMAN, R.L. Y DAVENPORT, J.M.: *Approximations of the critical region of the Friedman statistic, Communications in Statistics*, pp.571–595, 1980.

HAN, J. Y KAMBER, M.: D*ata mining: Concepts and techniques*, *Morgan Kauffmann Publishers, 2001*.

JOHN, G.H.: *Enhancements to the Data Mining Process,* Ph.D.Thesis, Stanford University, 1997.

LIPPMANN, R.: A*n introduction to computing with neural nets, BEE ASSP Magazine,* 1987.

MALERBA, D.: *Blocks Classification. Multistrategy Learning for Document Recognition Applied Artificial Intelligence*, pp.33-84, 1994.

MITCHELL, T.: *Machine Learning,* NY, McGraw Hill, 1997.

MOWFORTH, P. Y SHEPHERD, B.: V*ehicle silhouettes. Turing Institute Research Memorandum TIRM-87-018*, 1987.

NEMENYI, P.B.: *Distribution-free Multiple Comparisons,* PhD thesis, Princeton University, 1936.

PAO, Y.H.: *Adaptive Pattern Recognition and Neural Networks,* Ed. Addison-Wesley Reading, Massachusetts, 1989.

QUINLAN, J.: *C4.5: Programs for Machine Learning*, Morgan Kaufmann,1993 .

QUINLAN, J. R.: «Induction of Decision Trees (ID3 algorithm)», *Machine Learning,* Vol. 1**,** pp.81-106, 1986.

RICHARDS, J.A.: *Classification using neural networks, Remote Sensing Digital Image Analysis* Berlin, Springer-Verlag, 1993.

SCHLIMMER, J.: United States Congressional Voting Records Database. *In:* Congressional Quarterly Almanac, t. C., 2nd session 1984, Volume Xl, Washington, D.C., 1985.

SHESKIN, D.: *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, 2000.

SIGILLITO, V.: P*ima Indians Diabetes Database*, National Institute of Diabetes and Digestive and Kidney Diseases, 1990.

UCI: *Repository Of Machine Learning Databases* [Online]. http://www.ics.uci.edu/~mlearn/MLRepository.html. Visitado el 17 de abril de 2017.

Vol. 7, No. 2, abril-julio (2018)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

WEKA: *Waitako Environment for Knowledge Analysis*, 3.6.8 ed. University of Waikato Hamilton, New Zealand, 2012.

WILCOXON, F.: «Individual comparisons by ranking methods», *Biometrics,* Vol.1**,** pp.80–83, 1945.

ZAR, J. H.: *Biostatical Analysis, 1998*.