Una aplicación de conjuntos rugosos difusos en selección de características para la mejora de métodos de selección de instancias evolutivos

Joaquín Derrac[1], Salvador García[2], Chris Cornelis[3], and Francisco Herrera[1]

[1] Dept. de Ciencias de la Computación e Inteligencia artificial,
CITIC-UGR (Centro de Investigación en Tecnologías de la Información y las Comunicaciones).
Universidad de Granada, 18071 Granada, España.
jderrac@decsai.ugr.es,herrera@decsai.ugr.es
[2] Dept. de Ciencias de la Computación. Universidad de Jaén, 23071 Jaén, España.
sglopez@ujaen.es
[3] Dept. de Matemática Aplicada y Ciencias de la Computación. Universidad de Ghent, Gent,
Bélgica. chris.cornelis@ugent.be

Resumen Recientemente se han definido con éxito nuevos métodos de selección de características basados en la teoría de conjuntos rugosos difusos. Aunque por sí solos estos métodos permiten construir clasificadores de gran calidad, sus resultados pueden ser mejorados aun más si se emplean de forma conjunta con otras técnicas de preprocesamiento, como la selección de instancias.
En este trabajo presentamos un algoritmo híbrido para selección de instancias y características, orientado a mejorar la eficacia del clasificador del vecino más cercano. En él, proponemos el uso de un procedimiento de selección de características basado en conjuntos rugosos difusos junto a la búsqueda evolutiva realizada en el espacio de instancias. Los resultados obtenidos, contrastados teóricamente mediante técnicas estadísticas no paramétricas, muestran que nuestra propuesta obtiene una mejora de rendimiento significativa con respecto a las técnicas consideradas.

Keywords: Conjuntos Difusos Rugosos, Algoritmos Evolutivos, Selección de Instancias, Selección de Características, Clasificador del Vecino Más Cercano

1. Introducción

La reducción de datos es un proceso que puede aplicarse en situaciones en las que haya que analizar una gran cantidad de datos. Su objetivo consiste en seleccionar la información más representativa del conjunto de datos empleado. De esta manera, es posible mejorar los resultados de muchas aplicaciones de minería de datos, reduciendo su coste computacional y la necesidad de espacio de almacenamiento. Las técnicas de reducción de datos más conocidas son la Selección de Características (SC) [8], la Extracción de Características, la Discretización, la Generación de Instancias y la Selección de Instancias (SI) [7].

La Teoría de Conjuntos Rugosos (TCR) [9] se ha empleado recientemente para abordar la tarea de la SC. Esta técnica ha sido mejorada mediante el empleo de lógica difusa, obteniendo métodos que ofrecen una mayor flexibilidad y un mejor potencial a la

hora de seleccionar subconjuntos de características de gran calidad [3]. Por otro lado, en SI, los Algoritmos Evolutivos han emergido como una técnica de gran calidad, gracias a la posibilidad de definir la selección como un problema de búsqueda [5]. Ambos campos ofrecen herramientas apropiadas para la mejora de rendimiento de las técnicas de aprendizaje automático.

En este trabajo presentamos una nueva propuesta híbrida, que denominamos TCR-SIE (TCR aplicada a la SI Evolutiva). Nuestra propuesta emplea un Algoritmo Genético (AG) estacionario para seleccionar las instancias más prometedoras, mientras que las características son seleccionadas con un procedimiento heurístico difuso basado en TCR. Esta selección condicionará el funcionamiento del AG, modificando el entorno en que las instancias son seleccionadas.

Al acabar su ejecución, TCR-SIE reduce el conjunto de entrenamiento original mediante los mejores subconjuntos de instancias y características encontrados. Este conjunto estará listo para ser aplicado como conjunto de referencia para el clasificador del vecino más cercano (1-NN). Dicho conjunto es capaz de mejorar sustancialmente el rendimiento del clasificador, por encima del resto de técnicas consideradas de forma aislada, tal y como mostramos en el estudio experimental realizado (cuyos resultados han sido validados mediante el uso de técnicas estadísticas no paramétricas).

El resto del trabajo está organizado como sigue: La Sección 2 ofrece información preliminar sobre la SI evolutiva y la TCR difusos. La Sección 3 describe las principales características de TCR-SIE. La Sección 4 muestra el estudio experimental realizado y los resultados alcanzados. Finalmente, la Sección 5 resume nuestras conclusiones.

2. Preliminares

Esta sección se centra en describir dos temas: SI y SC como técnicas de reducción de datos (Sección 2.1), y la aplicación de la TCR difusos para SC (Sección 2.2).

2.1. Selección de instancias y características

El objetivo de la SI es aislar el conjunto de instancias más pequeño posible que permita funcionar a un algoritmo de minería de datos con la misma calidad que si empleara el conjunto de entrenamiento inicial [7]. Minimizando el conjunto de datos, el algoritmo de minería de datos ve reducido su coste computacional, tanto en tiempo como en espacio, y mejora su capacidad de generalización.

La SI se define como sigue: Sea $(\mathcal{X}, \mathcal{A})$ un sistema de decisión, donde $\mathcal{X} = \{x_1, \ldots, x_n\}$ y $\mathcal{A} = \{a_1, \ldots, a_m\}$ son conjuntos finitos no vacíos de instancias y características, respectivamente. Entonces, se asume la existencia de un conjunto de entrenamiento *TR* compuesto por *N* instancias, y un conjunto de test *TS* compuesto por *T* instancias ($TR \cup TS = (\mathcal{X}, \mathcal{A})$). Sea $S \subseteq TR$ el subconjunto de instancias seleccionado tras la aplicación de un algoritmo de SI. Así, toda instancia T de *TS* es clasificada por un algoritmo de minería de datos empleando tan solo las instancias contenidas en *S* como referencia.

Dentro del campo del aprendizaje automático se han desarrollado muchas propuestas evolutivas para SI [5, 6]. El interés en este campo creció con el estudio presentado por

Cano y otros [2]. En dicho estudio se concluye que los algoritmos evolutivos mejoran a los clásicos cuando se aplican a la SI, tanto en precisión de la etapa de clasificación como en el poder de reducción obtenido.

Por otro lado, la SC consiste en escoger aquellas características que mejor representen al conjunto de datos inicial. Así, es posible eliminar características redundantes e irrelevantes, para obtener clasificadores más simples y precisos [8]. De forma análoga a la SI, la SC puede describirse como sigue: Asumamos que $\mathcal{A}, \mathcal{X}$, *TR* y *TS* ya han sido definidos. Sea $B \subseteq \mathcal{A}$ el subconjunto de características seleccionadas por un algoritmo de SC actuando sobre *TR*. Así, toda instancia de *TS* es clasificada por un algoritmo de minería de datos empleando como referencia tan solo las características contenidas en B.

En la literatura se puede encontrar un gran número de propuestas para SC [8], incluyendo algunas recientes combinando SI y SC [4].

2.2. TCR difusa para SC

En el análisis de conjuntos rugosos [9], cada atributo a en $\mathcal{A}$ se identifica como una correspondencia $\mathcal{X} \rightarrow V_a$, en la que V_a es el conjunto de valores de a sobre $\mathcal{X}$. Para cada subconjunto B de $\mathcal{A}$, la B-indiscernible relación R_B se define como

$$R_B = \left\{(x, y) \in \mathcal{X}^2 \quad \text{y} \quad (\forall a \in B)(a(x) = a(y))\right\} \tag{1}$$

Por tanto, R_B es una relación de equivalencia. Sus clases $[x]_{R_B}$ pueden ser empleadas para aproximar conceptos, es decir, subconjuntos del universo $\mathcal{X}$. Dado $A \subseteq \mathcal{X}$, sus aproximaciones inferior y superior R_B se definen mediante

$$R_B \downarrow A = \{x \in \mathcal{X} | [x]_{R_B} \subseteq A\} \quad \text{and} \quad R_B \uparrow A = \{x \in \mathcal{X} | [x]_{R_B} \cap A \neq \varnothing\} \tag{2}$$

Un *sistema de decisión* $(\mathcal{X}, \mathcal{A} \cup \{d\})$ es un sistema de información especial, empleado en el contexto de clasificación, en el que d $(d \notin \mathcal{A})$ es un atributo denominado como *atributo de decisión*. Sus clases de equivalencia $[x]_{R_d}$ se denominan clases de decisión. Dada $B \subseteq A$, la región B-positiva POS_B contiene aquellos objetos de X para los que los valores de B permiten predecir la clase de decisión inequívocamente:

$$POS_B = \bigcup_{x \in X} R_B \downarrow [x]_{R_d} \tag{3}$$

Claramente, si $x \in POS_B$, se cumple que cuando una instancia tenga los mismos valores que x para los atributos de B, pertenecerá a la misma clase de decisión que x. La capacidad de predicción de los atributos de B con respecto a d se mide mediante la métrica γ (grado de dependencia de d en B):

$$\gamma_B = \frac{|POS_B|}{|\mathcal{X}|} \tag{4}$$

En lugar de utilizar una relación de equivalencia clásica para representar a R_B, podemos hacerlo mediante una relación difusa R. Típicamente, se asume que R es, al menos, una relación difusa de tolerancia (reflexiva y simétrica).

Asumiendo que se emplea el método clásico para discernir objetos para un atributo a, es decir, $R_a(x,y) = 1$ si $a(x) = a(y)$ y $R_a(x,y) = 0$ en otro caso, podemos definir la relación B-indiscernible difusa para cualquier subconjunto B de $\mathcal{A}$ como

$$R_B(x,y) = \mathcal{T}(R_a(x,y)), a \in B \tag{5}$$

en la que $\mathcal{T}$ es una t-norma. Se comprueba que si solo se usan atributos cualitativos, el concepto tradicional de B-indiscernibilidad permanece inalterado [3].

Para obtener los límites inferior y superior de un conjunto difuso A en X mediante una relación difusa de tolerancia R, reescribimos las fórmulas de (2) (empleando el implicador de Lukasiewicz $\mathcal{I}(x,y) = min(1, 1-x+y)$ y la t-norma mínimo $\mathcal{T}(x,y) = min(x,y), x, y \in [0,1]$) para definir $R \downarrow A$ y $R \uparrow A$, para todo y en X, como

$$(R \downarrow A)(Y) = \inf_{x \in X} \mathcal{I}(R(x,y), A(x)) \quad (R \uparrow A)(Y) = \sup_{x \in X} \mathcal{T}(R(x,y), A(x)) \tag{6}$$

Usando relaciones B-indiscernibles difusas, la región B-positiva difusa se define como

$$POS_B(y) = \left(\bigcup_{x \in X} R_B \downarrow [\mathcal{X}_{R_d}] \right)(y) \tag{7}$$

Una vez fijada la región positiva difusa, se puede definir una medida creciente valorada en el intervalo $[0,1]$ para medir el grado de dependencia de un conjunto de características sobre otro. Para la SC, es útil reescribir este concepto en términos del atributo de decisión:

$$\gamma_b = \frac{|POS_B|}{|POS_{\mathcal{A}}|} \tag{8}$$

3. TCR-SIE: TCR aplicada a la SI Evolutiva

Dedicamos esta parte a describir TCR-SIE. La Sección 3.1 describe el AG estacionario empleado para realizar la SI y el método de SC basado en TCR difusos. La Sección 3.2 muestra el modelo completo combinando ambas técnicas.

3.1. Técnicas básicas de TCR-SIE

La SI en TCR-SIE está guiada por un AG estacionario en que sólo se generan dos descendientes por generación. El resto de características importantes del AG comprenden codificación binaria, selección de padres mediante torneo binario, operador de cruce en 2 puntos y operador de mutación de cambio de bit. La función objetivo considera tanto mejorar el acierto en clasificación como reducir el tamaño del conjunto. Para ello, seguiremos la propuesta dada en [2], donde *Cano y otros* definieron *Pres* como la precisión obtenida por un clasificador 1-NN sobre el conjunto de entrenamiento completo, empleando el conjunto S actual como referencia y *leave-one-out* como esquema de validación; *Red* como el porcentaje actual de instancias descartadas, y un valor real, α,

para ajustar el peso de ambos términos en la función. La Ecuación 9 define la función objetivo, siendo J un cromosoma a evaluar

$$Fitness(J) = \alpha \cdot Pres(J) + (1 - \alpha) \cdot Red(J) \tag{9}$$

Siguiendo las recomendaciones dadas en [2], TCR-SIE empleará un valor $\alpha = 0{,}5$, para equilibrar adecuadamente ambos términos de la función.

El método de SC basado en TCR difusos se ha tomado de [3], donde se emplea una heurística clásica de *hillclimbing* (heurística *quickreduct*) para buscar subconjuntos de caracteristicas de forma iterativa, maximizando la medida γ (Ecuación 8). Para atributos numéricos, la medida de similaridad escogida es:

$$R_a(x,y) = max\left(min\left(\left(\frac{a(y) - a(x) + \sigma_a}{\sigma_a}, \frac{a(x) - a(y) + \sigma_a}{\sigma_a}\right), 0\right)\right) \tag{10}$$

donde x e y son dos instancias del conjunto de entrenamiento diferentes, y σ_a define la desviación estándar de a. Para atributos nominales, hemos empleado la métrica VDM [10], en la que dos valores se definen como cercanos si tienen una mayor correlación con los atributos de decisión.

3.2. Modelo híbrido para la aplicación simultánea de SI y SC

Una vez descritas las herramientas básicas para realizar SI y SC, es el momento de detallar la propuesta híbrida. Básicamente, se puede describir como un AG estacionario para SI donde, cada vez que un número fijo de evaluaciones ha sido gastado, se pone en marcha un proceso de SC basado en TCR difusos que altera las características que se tienen en cuenta durante la búsqueda.

1. **Inicialización**: Los cromosomas se inicializan aleatoriamente. Como conjunto inicial de características, se toma la mejor opción (en términos de precisión del clasificador 1-NN) entre dos posibles: El conjunto completo de características, o aquel devuelto por la aplicación del método de SC basado en TCR difusa sobre el conjunto de entrenamiento completo.
2. **Nueva generación de SI**: Cada generación de SI se aplica empleando el AG estacionario. Es importante destacar que, a la hora de evaluar un nuevo cromosoma, el clasificador 1-NN empleado en la función objetivo sólo tendrá en cuenta aquellas características actualmente seleccionadas por TCR-SIE.
3. **Actualizar características**: Si la fase de Estabilización no se ha activado aún (ver más abajo), se aplica un procedimiento de actualización de características seleccionada cada vez que se hayan gastado *ActualizarSC* evaluaciones. Este procedimiento consiste en aplicar la heurística *quickreduct* sobre el mejor cromosoma de la población, para obtener un nuevo subconjunto de características que lo represente. Si este nuevo subconjunto representa mejor al conjunto de entrenamiento original (es decir, ofrece mayor precisión al clasificador 1-NN que el anterior conjunto seleccionado por TCR-SIE), lo reemplaza durante el resto de la búsqueda.

4. **Fase de estabilización**: Los cambios en el conjunto de características seleccionado por TCR-SIE no son aceptados en la fase final del algoritmo. De esta manera, si el número de evaluaciones consumidas es mayor que $\beta \cdot$ *NEvaluaciones*, se activa la fase de Estabilización y no se permite actualizar el conjunto de características seleccionado durante el resto de la ejecución.
   Este mecanismo permite a TCR-SIE converger más fácilmente en problemas duros, gracias a que el conjunto final de características queda fijado antes del final de la búsqueda. Esto permite centrar los últimos esfuerzos de la búsqueda en refinar el conjunto final de instancias seleccionadas, una vez que el entorno de búsqueda ha quedado fijo.
5. **Criterio de parada**: El proceso de búsqueda termina cuando se hayan consumido *NEvaluaciones*. En otro caso, un nuevo ciclo del algoritmo comienza.
6. **Salida**: Cuando se han consumido *NEvaluaciones*, se extrae el mejor cromosoma de la población como conjunto final de instancias seleccionadas, y como conjunto final de características aquellas seleccionadas por TCR-SIE .

Los subconjuntos obtenidos por TCR-SIE definen una versión reducida del conjunto de entrenamiento original. Este nuevo conjunto puede ser usado como referencia por un clasificador 1-NN estándar, obteniendo resultados más precisos y siendo eficiente gracias a la reducción de tamaño obtenida.

4. Estudio experimental y resultados

Esta sección describe el estudio experimental realizado. Los conjuntos de datos, métodos de comparación y parámetros empleados se detallan en la Sección 4.1. Los resultados obtenidos se muestran en la Sección 4.2.

4.1. Estudio experimental

En nuestros experimentos, hemos usado 20 conjuntos de datos tomados del repositorio KEEL-Datasets [4] [1]. La Tabla 1 muestra sus principales características. Para cada conjunto, se detalla su número de instancias, características y atributos de decisión (clases). Todos ellos han sido empleados mediante un esquema de validación de 10 partes (*ten fold cross-validation*, (10-fcv))

Como métodos de comparación, hemos seleccionado aquellos considerados como técnicas básicas para la construcción de TCR-SIE (un AG estacionario para SI (AGE-SI) y el procedimiento de SC basado en TCR difusos (TCR-SC)). Los conjuntos de datos preprocesados obtenidos se han empleado como conjuntos de referencia para un clasificador 1-NN, para estimar su precisión. Además, hemos incluido el clasificador 1-NN empleando el conjunto original completo. La Tabla 2 muestra los parámetros empleados, cuyos valores han sido fijados de acuerdo a los recomendados en las propuestas anteriores del área.

Finalmente, emplearemos el conocido test de Wilcoxon para contrastar los resultados obtenidos. Para más información sobre éste test y otros procedimientos estadísticos

[4] http://www.keel.es/datasets.php

Conjunto	Instancias	Características	Clases	Conjunto	Instancias	Características	Clases
Australian	690	14	2	Housevotes	435	16	2
Balance	625	4	3	Iris	150	4	3
Bupa	345	6	2	Mammographic	961	5	2
Cleveland	303	13	5	Newthyroid	215	5	3
Contraceptive	1473	9	3	Pima	768	8	2
Ecoli	336	7	8	Sonar	208	60	2
German	1000	20	2	Tic-tac-toe	958	9	2
Glass	214	9	7	Wine	178	13	3
Hayes-roth	160	4	3	Wisconsin	699	9	2
Hepatitis	155	19	2	Zoo	101	16	7

Cuadro 1. Conjuntos empleados en el estudio experimental

Algoritmo	Parámetros
TCR-SIE	NEvaluaciones: 10000, Tam. Pob: 50, Prob. Cruce: 1.0, Prob. Mut.: 0.005 por bit, α: 0.5 *MaxGamma*: 1.0, ActualizarSC: 100, β: 0.75
AGE-SI	NEvaluaciones: 10000, Tam. Pob: 50, Prob. Cruce: 1.0, Prob. Mut.: 0.005 por bit, α: 0.5
TCR-SC	*MaxGamma*: 1.0
1-NN	-

Cuadro 2. Parámetros de los algoritmos empleados en el estudio experimental

específicamente diseñados para su uso en el área del aprendizaje automático, se puede visitar el sitio web temático del grupo SCI2S sobre *Inferencia Estadística en Inteligencia Computacional y Minería de Datos* [5].

4.2. Resultados obtenidos

La Tabla 3 muestra los resultados obtenidos en precisión (porcentaje de acierto en test), razón de reducción en instancias (Reducción (SI)) y razón de reducción en características (Reducción (SC)). Para cada conjunto, remarcamos en **negrita** el mejor resultado en precisión.

Como puede verse en la tabla, TCR-SIE obtiene la mayor precisión media en la fase de test. Para contrastar este resultado, hemos aplicado el test de Wilcoxon, cuyos resultados se muestran en la Tabla 4.

A partir de las Tablas 3 y 4, podemos realizar el siguiente análisis:

- En precisión, TCR-SIE obtiene el mejor resultado en 14 de 20 conjuntos, y el mejor resultado medio. Esta superioridad es identificada como significativa por el test de Wilcoxon, mostrando que TCR-SIE es superior al resto de métodos con un nivel de significancia $\alpha = 0{,}01$. Éste es un resultado fuerte, que indica que TCR-SIE mejora claramente al resto de alternativas, en términos de precisión.
- TCR-SIE obtiene resultados algo mejores que AGE-SI en términos de reducción sobre el conjunto de instancias. Por tanto, podemos afirmar que nuestra propuesta reduce de forma efectiva el conjunto de entrenamiento, a la par que mejora su

[5] http://sci2s.ugr.es/sicidm/

Medida	Precisión				Reducción (SI)		Reducción (SC)	
Conjunto	TCR-SIE	AGE-SI	TCR-SC	1-NN	TCR-SIE	AGE-SI	TCR-SIE	TCR-SC
Australian	**85.66**	85.65	81.45	81.45	0.8872	0.8799	0.1571	0.0000
Balance	85.92	**86.40**	79.04	79.04	0.8464	0.8686	0.0000	0.0000
Bupa	**65.72**	61.14	62.51	61.08	0.8502	0.8644	0.0000	0.1274
Cleveland	**55.16**	52.82	52.51	53.14	0.9014	0.9171	0.0462	0.3908
Contraceptive	**45.42**	44.54	42.63	42.77	0.7637	0.7530	0.0667	0.0360
Ecoli	**82.14**	80.38	76.58	80.70	0.8882	0.9077	0.1286	0.2286
German	**70.80**	70.40	67.90	70.50	0.8014	0.7914	0.2350	0.1450
Glass	67.35	67.10	**74.50**	73.61	0.8718	0.8791	0.0444	0.0168
Hayes-roth	**80.86**	69.15	76.07	35.70	0.8544	0.8384	0.2500	0.1000
Hepatitis	**82.58**	79.33	79.50	82.04	0.9262	0.9226	0.5368	0.4263
Housevotes	**94.48**	93.79	90.78	91.24	0.9387	0.9410	0.3500	0.0188
Iris	**96.00**	94.67	93.33	93.33	0.9511	0.9481	0.1250	0.0000
Mammographic	**80.65**	79.50	75.76	76.38	0.8322	0.8229	0.0000	0.3396
Newthyroid	96.77	**98.16**	97.23	97.23	0.9473	0.9571	0.0600	0.0000
Pima	**74.80**	72.26	70.33	70.33	0.7911	0.8187	0.0000	0.0000
Sonar	80.76	75.45	81.69	**85.55**	0.8899	0.8595	0.2900	0.7183
Tic-tac-toe	78.29	**78.71**	73.07	73.07	0.8655	0.7917	0.0000	0.0000
Wine	**97.19**	92.68	95.49	95.52	0.9451	0.9538	0.3308	0.5231
Wisconsin	**96.42**	96.13	95.57	95.57	0.9103	0.9027	0.0444	0.0000
Zoo	96.39	94.22	**96.50**	92.81	0.8634	0.8714	0.2125	0.2750
Media	**80.67**	78.63	78.12	76.55	0.8763	0.8745	0.1439	0.1673

Cuadro 3. Resultados obtenidos

Comparación	R^+	R^-	P-value
TCR-SIE vs AGE-SI	188	22	0.0010
TCR-SIE vs TCR-SC	183	27	0.0023
TCR-SIE vs 1-NN	174	36	0.0083

Cuadro 4. Resultados del test de Wilcoxon

eficacia. En el espacio de características, obtiene una reducción similar a la de TCR-SC, si bien las características seleccionadas son diferentes, en general.

Estos resultados confirman los beneficios de hibridar los métodos de SI evolutivo y SC basada en TCR difusos en una única propuesta. Además, destacan a TCR-SIE como una propuesta de preprocesamiento apropiada para la reducción del tamaño del conjunto de entrenamiento (en más de un 87 %, en media) a la par que mejora la precisión del clasificador 1-NN.

Para finalizar el estudio, la Figura 1 representa la comparación de resultados entre TCR-SIE y el clasificador base sin aplicar preprocesamiento (1-NN). A cada conjunto se le asigna un punto, donde su valor en el eje de abscisas indica la precisión obtenida por TCR-SIE, mientras que su valor en el eje de ordenadas indica la precisión obtenida por el clasificador 1-NN. La figura muestra claramente la mejora obtenida, como puede apreciarse al ver que la mayoría de los puntos (17 de 20) quedan por debajo de la línea diagonal que corta la gráfica (que representa una precisión similar entre ambos métodos).

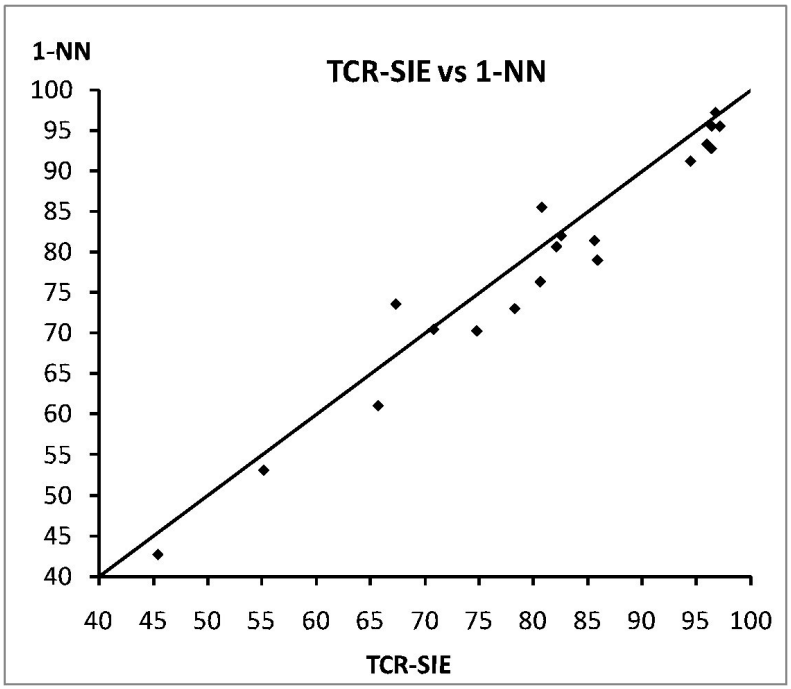

Figura 1. Representación de TCR-SIE vs 1-NN. Puede apreciarse que la mayoría de los puntos quedan por debajo de la diagonal (igualdad en rendimiento), destacando, por tanto, la mejora obtenida tras la aplicación del proceso de preprocesamiento

5. Conclusiones

En este trabajo hemos presentado TCR-SIE, una nueva propuesta que integra mecanismos de SI evolutiva y SC basada en la TCR difusa. Esta propuesta incluye las características seleccionadas por el método de TCR difuso dentro de la búsqueda evolutiva, combinando los beneficios de ambas técnicas en un único y preciso procedimiento.

Los resultados obtenidos muestran que nuestra propuesta obtiene una mejor precisión que las técnicas consideradas de forma aislada, manteniendo una capacidad de reducción del conjunto de entrenamiento similar. Los procedimientos estadísticos no paramétricos empleados confirman que podemos considerar a TCR-SIE como una herramienta apropiada para la mejora del clasificador 1-NN.

Como trabajo futuro, se plantea ampliar el estudio experimental con un conjunto más amplio de técnicas del estado del arte, así como considerar la aplicación de TCR-SIE sobre nuevos tipos de clasificadores, distintos del 1-NN.

Agradecimientos

Este trabajo ha sido soportado por los proyectos TIN2011-28488 y P10-TIC-6858. J. Derrac posee una beca FPU del Ministerio de Educación.

Referencias

1. Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., Herrera, F.: Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. Journal of Multiple-Valued Logic and Soft Computing 17(2-3) (2011)
2. Cano, J.R., Herrera, F., Lozano, M.: Using evolutionary algorithms as instance selection for data reduction in KDD: An experimental study. IEEE Transactions on Evolutionary Computation 7(6), 561–575 (2003)
3. Cornelis, C., Jensen, R., Hurtado, G., Slezak, D.: Attribute selection with fuzzy decision reducts. Information Sciences 180, 209–224 (2010)
4. Derrac, J., García, S., Herrera, F.: IFS-CoCo: Instance and feature selection based on cooperative coevolution with nearest neighbor rule. Pattern Recognition 43(6), 2082–2105 (2010)
5. Derrac, J., García, S., Herrera, F.: A survey on evolutionary instance selection and generation. International Journal of Applied Metaheuristic Computing 1(1), 60–92 (2010)
6. García, S., Derrac, J., Cano, J.R., Herrera, F.: Prototype selection for nearest neighbor classification: Taxonomy and empirical study. IEEE Transactions on Pattern Analysis and Machine Intelligence, in press (2011)
7. Liu, H., Motoda, H. (eds.): Instance Selection and Construction for Data Mining. The Springer International Series in Engineering and Computer Science, Springer (2001)
8. Liu, H., Motoda, H. (eds.): Computational Methods of Feature Selection. Chapman & Hall/Crc Data Mining and Knowledge Discovery Series, Chapman & Hall/Crc (2007)
9. Pawlak, Z., Skowron, A.: Rudiments of rough sets. Information Sciences 177, 3–27 (2007)
10. Wilson, D., Martinez, T.: Improved heterogeneous distance functions. Journal of Artificial Intelligence Research 6, 1–34 (1997)