Modelos de predicción de actividad citotóxica en células SK-N-SH mediante técnicas de *softcomputing* en una muestra heterogénea de compuestos

Julio Omar Prieto-Entenza, Mario Pupo-Merino* y Ramón Carrasco-Velar

Universidad de las Ciencias Informáticas, Carretera a San Antonio de los Baños, Kilómetro 2½, Boyeros, La Habana, Cuba.
*Universidad Central de Las Villas.

Recibido: 7 de junio de 2010. Aceptado: 30 de marzo de 2011.

Palabras clave: citotoxicidad, muestras desproporcionadas, fragmentos, índice del estado refractotopológico, estructura-actividad.
Key words: citotoxicity, unbalanced data set, fragment, refractotopological state index, Structure-activity.

RESUMEN. Las técnicas de *softcomputing* consisten en la combinación de técnicas de inteligencia artificial para la solución de problemas. En los estudios de relación entre la estructura química y la actividad biológica se han venido aplicando con éxito. En el presente trabajo se expone la aplicación de estas técnicas para el desarrollo de modelos de relación estructura -actividad en una muestra heterogénea de 1 335 compuestos evaluados en el ensayo sobre viabilidad celular en células SK-N-SH para determinar *in vitro* la citotoxicidad de pequeñas moléculas (Ensayo 435 de la base de datos del NCBI). Se aplicó un sistema de inferencia borroso evolutivo a dicha muestra y se compararon los diferentes algoritmos aplicados. La muestra seleccionada presentó una amplia diversidad estructural y una relación entre las clases activa e inactiva de 1 : 12, lo cual constituyó un gran desbalance entre ellas. En el pre-procesamiento de la muestra se emplearon los métodos SMOTE e hibrido que contribuyen a reducir su desbalance hasta uniformarla. En el trabajo se definió además, el Índice del Estado Refractotopológico Total de un fragmento como la suma de los valores del índice correspondiente de los átomos pesados que lo conforman, el cual se empleó como descriptor estructural. Se logró una clasificación correcta de hasta 70,6 %, en dependencia del algoritmo utilizado, lo cual constituyó un resultado aceptable, teniendo en cuenta la elevada diversidad estructural de la muestra y la heterogeneidad de las fuentes de datos biológicos.

ABSTRACT. The soft computing techniques are a combination of different artificial intelligent algorithms employed for the solution of different problems that had been broadly employed in structure activity relationships studies. In this work, the application of these techniques for developing of structure-activity models in a heterogeneous set of 1 335 compounds evaluated in the cellular viability assay with SK-NH-SH cells to determine cytotoxicity of small molecules was done (assay No. 435 in the NCBI Database). A fuzzy evolutionary inference system was applied and the different algorithms were compared. The selected sample showed a great structural diversity and the relationship between active and inactive classes was 1 : 12 which was considered a great unbalanced data set. The sample was preprocessed using both SMOTE and hybrid methods to reduce this handicap. In this work was also defined the Total Refractotopological State Index for Fragments as the sum of the corresponding index for each one of the atoms included in the fragment, which was employed as structural descriptor. A correct classification of 70.6 % was obtained, in dependence of the employed algorithm. This is an acceptable result, taking into account the high structural diversity of the compounds set and the heterogeneity of the data source.

INTRODUCCIÓN

Desde un punto de vista formal, los estudios de relación estructura química-actividad biológica constituyen el procedimiento para establecer, mediante métodos matemáticos, el vínculo que existe entre las características estructurales de las moléculas y sus propiedades biológicas determinadas experimentalmente.[1] Para las primeras se utilizan, desde mediciones químico físicas hasta descriptores teóricos y para las segundas, ensayos biológicos realizados *in vitro* o *in vivo*. Todos esos datos son procesados posteriormente con diferentes herramientas matemáticas que establecen los modelos correspondientes.

El problema en la clasificación de bases de datos no balanceadas es que los algoritmos de clasificación tienen tendencia a describir la clase mayoritaria. Esto ocurre

Correspondencia:
Dr. Ramón Carrasco Velar
Universidad de las Ciencias Informáticas, Carretera a San Antonio de los Baños, kilómetro 2½, Boyeros, La Habana, Cuba. Correo electrónico: rcarrasco@uci.cu

puesto que el clasificador intenta reducir el error global y este no tiene en cuenta la distribución de los datos.

Variables descriptoras y el enfoque fragmental

Una forma de establecer correlaciones entre la estructura química y la actividad biológica es describiendo la molécula como una colección de fragmentos derivados de ella.[2] Las moléculas no necesariamente presentan su actividad biológica como un ente único sino que una parte de ellas es la responsable directa de la actividad. Este es el caso, por ejemplo, de los antibióticos del tipo de los beta-lactámicos, para los cuales se ha empleado el procedimiento de segmentar la molécula en partes de interés. Este enfoque fragmentario puede verse desde los trabajos iniciales de Hammet[3] en 1937 para la definición de las constantes electrónicas de grupos sustituyentes. Más hacia la actualidad y basados en la teoría del grafo químico, se han desarrollado múltiples descriptores topológicos grafo-teóricos como los de Wiener, Randic y Kier y Hall,[4,5,6] los topográficos[7-11] y más recientemente, los híbridos.[12,13] Los índices topológicos son conceptualmente simples y de fácil interpretación, ya que emplean información basada solamente en la conexión entre los átomos de la molécula y las aplicaciones informáticas que los calculan son computacionalmente eficientes.[14] Los topográficos incluyen información estructural 3D a partir de cálculos semiempíricos y los híbridos incorporan información sobre alguna propiedad químico física particionada sobre los átomos. Sin embargo, se ha demostrado por Estrada y cols.,[11] que los índices topológicos explican más del 70 % del comportamiento de la estructura y se argumenta que su simplicidad de cálculo los convierte en la opción de elección. No obstante, la amplia diversidad estructural de la muestra empleada sugiere la necesidad de compensar esta dificultad con un incremento en el contenido de información de los descriptores para minimizar la posible presencia de *outliers* como consecuencia de pertenecer a diferentes espacios de parámetros. En relación con este aspecto, Kühne y cols.[15] desarrollaron una metodología para caracterizar el dominio químico de modelos SAR y QSAR, basada en un enfoque de fragmentos centrados en átomos (*Atom Centered Fragment, ACF)* en el cual descomponen la molécula en piezas del grafo desprovisto de hidrógeno que actúan como centros ACF cuyos respectivos tamaños varían en dependencia del camino hacia los átomos vecinos incluidos los átomos de hidrógeno.

Las técnicas de *softcomputing*

En la creación de modelos de predicción mediante descriptores grafo-teóricos, se han utilizado varias técnicas de inteligencia artificial para el procesamiento de los datos. Se ha destacado recientemente, el empleo de técnicas de *softcomputing*. Este término lo introdujo Zadeh[16] para denotar una aproximación al razonamiento humano que deliberadamente hace uso de imprecisiones y vaguedades para obtener soluciones razonables que son fáciles de manipular. Bajo este principio, los sistemas borrosos o difusos, las redes neuronales (RN), la computación evolutiva, el razonamiento probabilístico y las combinaciones de dichos métodos se consideran como *softcomputing*. Estas técnicas no tienen por lo general una propuesta de solución única de un problema; sin embargo, son robustas ante entornos con entradas ruidosas, tienen una gran tolerancia a la imprecisión de los datos con los que se trabaja y pueden ser explotadas para ganar robustez con soluciones de bajo costo y mayor capacidad de modelación.[17]

En el desarrollo de técnicas basadas en *softcomputing,* desempeña un papel importante la generación de reglas borrosas de la forma *If-then* creando los denominados sistemas de inferencia borrosos (SIB). Actualmente existen varios enfoques para el aprendizaje y la optimización de las reglas. El más empleado consiste en la creación inicial de una base de reglas borrosas o difusas a partir de los datos utilizados en el aprendizaje y su posterior optimización.[18] En la estrategia basada en dicha optimización, se perciben dos líneas fundamentales: la aplicación de algoritmos derivados de redes neuronales que emplean para su aprendizaje elementos basados en el gradiente o heurísticas[19,20] y el uso de los algoritmos evolutivos (AE).[21] El uso de las RN ha mostrado una gran eficiencia en la resolución de diversos problemas, sin embargo, se les señala que se comportan como cajas negras y que no facilitan la interpretación de los resultados.[22] En el caso de los algoritmos evolutivos, hay que destacar su facilidad de aplicación y las pocas restricciones que imponen a los problemas. Se aprecia en este sentido, el uso de los algoritmos genéticos (AG).[23] Su principal ventaja con respecto a los sistemas neuroborrosos es su capacidad de presentar en un espacio de soluciones la integración de variables de diferente naturaleza como binarias, discretas o reales, con la definición de operadores específicos que permiten la evolución de estas estructuras. Esto permite combinar mecanismos de selección de reglas con el ajuste de parámetros y el aprendizaje de funciones de pertenencia representadas con variables discretas, junto con la base de reglas. De esta forma, se logran modelos con una mejor capacidad de interpretación de los resultados.[22,24]

Clases desproporcionadas

La aplicación de técnicas de aprendizaje computacional a problemas reales ha traído consigo una serie de retos que previamente no habían sido considerados como relevantes. Entre estos problemas se encuentran, el ruido en los datos, así como el solapamiento y el desbalance entre clases. El primero se deriva del gran parecido que pueden presentar los datos pertenecientes a clases diferentes o errores en los datos. El segundo se presenta cuando datos pertenecientes a clases diferentes ocupan un espacio común porque algunos de los atributos de ellas comparten el mismo rango de valores. El tercer problema es el de las clases desproporcionadas, que tiene lugar cuando se poseen muchos ejemplos de una clase, pero relativamente muy pocos de otra.

En los problemas de clasificación reales ocurre con frecuencia que las clases no están equilibradas, de forma que aparecen muchos más ejemplos de una clase de que otra y suele ocurrir que la clase minoritaria es la clase de interés. La mayoría de los métodos de clasificación para obtener una buena precisión, se centran en la clase mayoritaria. Entre las acciones que se realizan para el tratamiento del ruido se encuentran, el empleo de algoritmos tolerantes al ruido y la utilización de algoritmos que filtren las instancias ruidosas o algoritmos que corrijan dichas instancias ruidosas. Cuando en una muestra se presenta un solapamiento, los datos pueden *limpiarse* con el empleo de algoritmos diseñados al efecto. Para el caso de existencia de desbalanceo en una muestra, se utilizan diferentes procedimientos:[25,26]

■ ***Oversampling*** *o sobre muestreo*: Consiste en balancear la distribución de las clases añadiendo ejemplos de la clase minoritaria. Algunos de los algoritmos más representativos son:

◆ *SMOTE*[27] (*Synthetic Minority Oversampling TEchnique*) que genera nuevas instancias de la clase minoritaria interpolando los valores de las instancias minoritarias más cercanas a una dada y

◆ *Resampling*. que duplica al azar instancias de la clase minoritaria.

■ ***Undersampling*** *o submuestreo.* Consiste en eliminar ejemplos de la clase mayoritaria.

◆ *Random undersampling*: Elimina al azar instancias de la clase mayoritaria.

◆ *Tomek Links* o enlaces de Tomek[27,28]: Elimina sólo instancias de la clase mayoritaria que sean redundantes o que se encuentren muy cerca de instancias de la clase minoritaria.

◆ *Wilson Editing*. También conocido como ENN (Editing Nearest Neighbor) elimina aquellas instancias donde la mayoría de sus vecinos pertenecen a otra clase.

■ ***Boosting.*** Esta técnica consiste en asociar pesos a cada instancia que se va modificando en cada iteración del clasificador. Al inicio, todas las instancias tienen el mismo peso y después de cada iteración, en función del error cometido en la clasificación, se reajustan los pesos con el objetivo de reducir dicho error. Dentro de esta técnica se encuentra *AdaBoost (boosting adaptativo)*, que implementa el algoritmo de *Boosting* descrito. En cada iteración, *AdaBoost* genera nuevas instancias mediante el empleo del algoritmo *Resampling*.

Todas estas técnicas tienen sus ventajas y sus inconvenientes. Entre estos en el caso de *undersampling*, se tiene la pérdida de información que se produce al eliminar instancias de la muestra. Sin embargo, tiene la ventaja de que reduce el tiempo de procesamiento del conjunto de datos. *Oversampling* tiene la ventaja de no perder información, pero puede repetir muestras con ruido, además de aumentar el tiempo necesario para procesar el conjunto de datos. Las técnicas *boosting* tienden a equilibrar estas deficiencias de los métodos de *oversampling* y *undersampling*.

El tema de la desproporción entre clases ha cobrado recientemente gran interés en la comunidad científica, debido a que es un problema relativamente común en una gran cantidad de situaciones reales, y porque los algoritmos actuales de aprendizaje tienen pobres desempeños para la clase minoritaria, la cual por lo general, es justamente la que interesa clasificar correctamente.[29]

Este tipo de problema se ha abordado con diferentes propuestas tales como, nuevas medidas para evaluar el desempeño de clasificadores ante clases desbalanceadas, así como soluciones a nivel de los datos y a nivel algorítmico.[30]

Los grafos ROC (*Receiver Operating Characteristic*) se utilizan para analizar la relación entre verdaderos positivos y negativos. En 1997, Bradley[31] sugirió el empleo del área bajo la curva ROC (*ABC*, Ec. 1) para brindar un valor sencillo del comportamiento de la clasificación. Cuanto mayor es el área, mejor es la clasificación y su valor máximo es 1, ya que corresponde al área de un cuadrado con dimensiones 1x1. Riquelme *et. al.*[32] demostraron por su parte que, aunque las técnicas de balanceo no mejoran el porcentaje de instancias clasificadas correctamente, la medida de *ABC* aumenta, es decir, se clasifican mejor las instancias de la clase minoritaria, que como ya se ha planteado, son normalmente las de interés.

$$ABC = \frac{1 + VP_{rate} - FP_{rate}}{2} \quad (1)$$

donde:

VP_{rate} Fracción de los casos positivos correctamente clasificados como pertenecientes a la clase *positivos*.

FP_{rate} la fracción de casos negativos incorrectamente clasificados como pertenecientes a la clase *positivos*.

Como medida de clasificación en el presente trabajo se emplea el área bajo la curva *ABC*.

Una estrategia de submuestreo es eliminar solo ejemplos de la clase mayoritaria que sean redundantes o muy alejados de la clase minoritaria, mediante lo que se conoce como *enlaces de Tomek o Tomek links*,[28] que elimina solo instancias de la clase mayoritaria que sean redundantes o que se encuentren muy cerca de instancias de la clase minoritaria. CNN (*Condensed nearest neighbour rule*)[33] que pretende eliminar los ejemplos de la clase mayoritaria que están distantes de la frontera de decisión basado en la construcción, paso a paso, de un nuevo conjunto de objetos V, moviendo hacia él cada elemento de la matriz de entrenamiento T, si este es erróneamente clasificado por los objetos que ya están en V. o *One-sided selection* (OSS)[34] que es un método de selección de instancias resultado de la aplicación de los enlaces de Tomek, seguido de la aplicación de CNN.

Por otro lado, uno de los métodos más utilizados de sobre-muestreo, el SMOTE[35], es un algoritmo de *oversampling* que genera instancias "sintéticas" o artificiales para equilibrar la muestra de datos basado en la regla del vecino más cercano. La generación se realiza extrapolando nuevas instancias en lugar de duplicarlas como hace el algoritmo de *Resampling*.

Genera nuevos ejemplos de la clase minoritaria a partir de los casos existentes a través de la interpolación de los valores resultantes de vecinos más cercanos entre ejemplos de la clase minoritaria. Este es el procedimiento más utilizado para realizar el sobremuestreo, pero tiene como principal riesgo que la generación de muchos nuevos casos puede ser causa de ruido en los datos que ocasione alteración de los resultados. El método SMOTE funciona solo con variables continuas y resulta más efectivo mientras más diferenciados estén entre sí los miembros de la clase minoritaria.

Recientemente, el grupo de *softcomputing* de la Universidad de Granada ha comenzado a realizar estudios detallados sobre este tema con el empleo de técnicas de preprocesamiento y varios algoritmos difusos.[36,37] Han demostrado que en general, los resultados de la clasificación mejoran con el empleo del método SMOTE como técnica de preprocesamiento y el uso de sistemas de inferencia difusos basados en reglas con un enfoque lingüístico.[38]

El Índice del Estado Refractotopológico para Átomos

Este índice híbrido[12] se inspiró en su homólogo, el Índice del Estado Electrotopológico para Átomos[39] de Kier y Hall y, a diferencia de este, se definió a partir de la matriz de adyacencia del grafo molecular completo, ponderado en los vértices por el valor de la refractividad atómica calculada por el método de Ghose y Crippen.[40,41] El índice brinda una idea de cuán sumergido o bloqueado se encuentra el átomo que se analice, según el valor que adquiere en su entorno molecular.[1,12,42] Este índice ha sido utilizado también para el mapeo de grupos farmacofóricos[43] por su capacidad de revelar el efecto puntual de las fuerzas de dispersión de London en la actividad biológica. Por otra parte, en su evaluación, se demostró que la suma de los valores del índice para cada uno de los átomos presentes, es igual a la refractividad molecular del compuesto.[1] De esto último, se infiere que, si la suma se realiza sobre los átomos pesados de un fragmento cualquiera, ese valor total corresponderá a la refractividad del fragmento específico.

El objetivo del presente trabajo consistió en desarrollar modelos de clasificación en una muestra heterogénea de compuestos evaluados por su acción citotóxica frente a células SK-N-SH con el empleo de técnicas de *softcomputing* y la aplicación de una variante del Índice del Estado Refractotopológico para átomos.

MATERIALES Y MÉTODOS

Datos biológicos

Para este trabajo, se seleccionó el ensayo sobre la viabilidad celular en células SK-N-SH para determinar *in vitro* la citotoxicidad de pequeñas moléculas[44] (Ensayo 435 de la base de datos Pubchem bioassay). En este ensayo, se reportan un total de 1 335 compuestos evaluados los cuales presentan una gran diversidad estructural [Fig. 1 (A-D)].

Preprocesamiento de la muestra

En los datos originales 55 compuestos presentaban resultados no concluyentes de su actividad. Como activos se reportaban 86 y el resto como inactivos. No obstante, se decidió limitar las clases de compuestos solamente a dos identificadas como activos-inactivos. Para realizar este reordenamiento de la muestra, se seleccionó el algoritmo XMeans, debido a su rapidez y a la posibilidad que brinda para definir las cantidades mínima y máxima de grupos que debe formar. Como resultado, se crearon dos nuevos grupos en los que 101 casos correspondieron a compuestos considerados como activos y 1 225 inactivos. Los pertenecientes a la clase de los noconcluyentes se distribuyeron entre las dos nuevas clases. A la relación entre la clase mayoritaria y la minoritaria se le denominó índice de desbalance (*IR*) y tuvo en la muestra empleada un valor final de $IR = 12{,}1$. Se consideró que valores de $IR > 10$ calificaban a una muestra como altamente desproporcionada.[45]

La nueva muestra se sometió entonces a un segundo proceso. En una primera variante, se aplicó el método de sobredimensionamiento SMOTE con valores de vecindad $k = 1$ y $k = 5$ para obtener los grupos de compuestos designados como FSM1 y FSM5. En una segunda variante se aplicó el método híbrido, compuesto por la unión del método de sobredimensionamiento SMOTE seguido del empleo del algoritmo de submuestreo CNN, lo que dio lugar a otros dos grupos de compuestos, los cuales se denominaron FSM1CNN y FSM5CNN.

Herramientas empleadas

En el procesamiento de los datos, se utilizó el paquete de programas KEEL[46] V.1.3 (Knowledge Extraction based on Evolutionary Learning), el cual permite evaluar algoritmos evolutivos en problemas de minería de datos y que incluye experimentos de regresión, clasificación y aprendizaje no supervisado, así como su integración con técnicas de preprocesamiento de muestras, lo que posibilita realizar un análisis completo de cualquier modelo de aprendizaje.

Los algoritmos seleccionados para este trabajo fueron Chi-RW,[47] SGERD (Algoritmo genético de estado estacionario para la extracción de reglas difusas de clasificación a partir de los datos),[48] AdaBoost,[49] LogiBoost[50] y Max-Logiboost[51] para sistemas de inferencia borrosos.

Para la validación de los resultados, se empleó el paquete estadístico SPSS V.13.1[52]

Se aplicaron varias pruebas no paramétricas para determinar la calidad de la clasificación y el comportamiento de los algoritmos difusos empleados. Los resultados de la clasificación se computaron a través de los valores del área bajo la curva *ROC* de cada muestra.

RESULTADOS Y DISCUSIÓN

Índice del Estado Refractotopológico Total para Fragmentos

Como fue referido, la suma de los valores de todos los átomos pesados de la molécula aislada, corresponde a su refractividad molecular MR. Se deduce entonces que, si la suma de los valores de $\Re$ en la molécula corresponde a su valor de MR, la suma de los valores de $\Re$ de los átomos de un fragmento dado corresponderán a la refractividad de dicho fragmento, con la característica particular de que esa refractividad corresponderá a la que se derive de la influencia de *su entorno molecular*. Debe tenerse en cuenta que el cálculo de la refractividad molecular de una molécula se realiza a partir de la molécula aislada. De esta manera, será posible distinguir fragmentos topológicamente iguales por su capacidad de adquirir un valor diferente en dependencia de los respectivos entornos moleculares. Se define entonces el Índice del Estado Refractotopológico Total para Fragmentos como:

$$\Re_T = \sum_i^j \Re_i \qquad (2)$$

donde:

$\Re_T$ Índice del Estado Refractotopológico Total del Fragmento.

$\Re_i$ Valor correspondiente al átomo i del fragmento analizado. La sumatoria es sobre todos los átomos pesados que componen dicho fragmento.

La (Tabla 1) presenta una serie de ejemplos de fragmentos topológicamente iguales que corroboran la aseveración anterior. Se seleccionó a la xantina y la hipoxantina atendiendo a la elevada similitud estructural que presentan. En estos compuestos puede observarse como no obstante las pequeñas diferencias, el índice es capaz de distinguir fragmentos topológicamente iguales

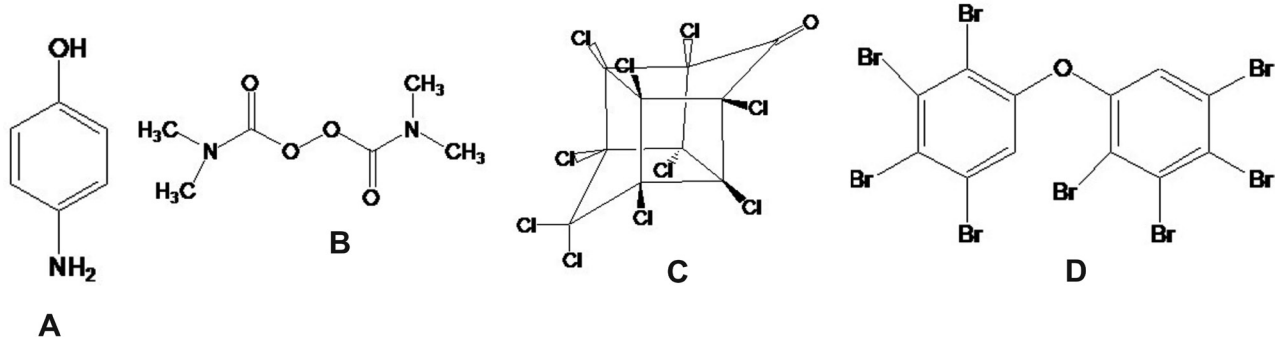
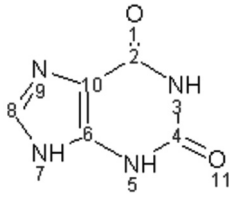
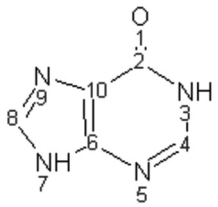


Fig. 1. *Estructuras de algunos de los compuestos evaluados A) 4-Aminofenol. B) Thiram. C) Clordecona. D) Octabromodifenilo.*

Tabla 1. Ejemplo de fragmentos topológicamente iguales con valores diferentes de $\Re_T$.

Xantina		Hipoxantina	
			
Fragmento	$\Re_T$	Fragmento	$\Re_T$
1 – 2 (11 – 4)	3,693 1 (3,581 1)	1 – 2	3,297 9
1 – 2 – 3 (11 – 4 – 3 y 11 – 4 – 5)	8,697 7 (8,585 7; 8,432 4)	1 – 2 – 3	7,797 0
6 – 10	5,540 8	6 – 10	7,202 9
8 – 9	7,693 1	8 – 9	7,311 4
7 – 8	9,243 2	7 – 8	9,220 6
1 – 2 – 10	6,568 5	1 – 2 – 10	7,056 5
3 – 4 – 5	12,837 0	3 – 4 – 5	12,973 3
7 – 8 – 9	12,256 6	7 – 8 – 9	11,039 9
2 – 3 – 4	11,094 9	2 – 3 – 4	11,769 8
6 – 7 – 8 – 9 – 10	17,797 4	6 – 7 – 8 – 9 – 10	18,242 7

$\Re_T$ Índice del Estado Refractotopológico Total de Fragmentos.

por su diferente capacidad de revelar las fuerzas dispersivas de London representadas por el Índice del Estado Refractotopológico Total de Fragmentos.

Un análisis (Tabla 1) permite corroborar que, fragmentos que son estructural y topológicamente iguales, son diferenciables por el valor correspondiente al Índice del Estado Refractotopológico Total para Fragmentos. Así, es posible observar que fragmentos idénticos desde el punto de vista topológico y que se encuentran en la misma molécula, son distinguibles por el valor que adquiere este índice en dependencia de sus correspondientes alrededores químicos. También se puede observar la diferencia con los fragmentos similares en la estructura análoga. Según estos resultados, la hipoxantina debe presentar una capacidad ligeramente mayor para la formación de dipolos instantáneos que su análogo xantina. Por un razonamiento similar, se puede afirmar que el doble enlace entre los átomos de carbono 6 y 10 es más polarizable en la hipoxantina que en la xantina. El cálculo de las cargas netas de los fragmentos en las correspondientes estructuras optimizadas por el método semiempírico PM3[53] con el programa Mopac[54] las estimó en - 0,474 y - 0,415 para la xantina y la hipoxantina, respectivamente. Esto revela una elevada carga formal sobre el doble enlace interno, mientras que los valores calculados de $\Re_T$ de 5,540 8 y 7,202 9 indican que en el caso de la hipoxantina, su doble enlace común es más polarizable.

Modelos

Para el desarrollo de los modelos se emplearon como fragmentos tipo los caminos de orden 2 y 3; ciclos desde orden 3 a 9, *clusters* de orden 3 y 4, y combinaciones clusters-caminos para caminos 2 y 3. Estos tipos de subgrafos se seleccionaron siguiendo como criterio, que este tipo de fragmentos son los más abundantes y que como regla, coinciden con grupos funcionales presentes en casi cualquier molécula, así como para evitar la generación de un volumen muy grande de datos que no aporta información útil por la gran cantidad de valores nulos determinados por la elevada diversidad estructural, lo cual sería el caso si se incluyeran cadenas de más de cuatro átomos, (Tabla 2). La (Tabla 3) resume las características de los conjuntos de datos generados. El algoritmo generó una cifra casi igual de inactivos-activos (1 254 inactivos; 1 260 activos; IR = 0,995). En las combinaciones SMOTE-CNN la relación inactivos-activos resultó de 339-322 y 439-418 para k = 1 y k = 5 respectivamente. Para estas últimas muestras, los valores correspondientes de IR fueron 1,02 y 1,05 respectivamente. En todos los casos, se utilizaron las 14 variables independientes correspondientes a los diferentes fragmentos.

A partir de las cuatro nuevas muestras, se conformaron las series de entrenamiento y prueba (cinco por cada una), cuyos resultados se promediaron. A cada muestra de entrenamiento se le aplicaron los diferentes algoritmos borrosos y se realizó la predicción en las correspondientes muestras de prueba. (Tabla 4).

Las medias las predicciones se encontraron en un rango relativamente bajo comprendido entre 0,54 y 0,65, no obstante lo cual constituye en principio, un resultado aceptable teniendo en cuenta la gran diversidad estructural, ya que siempre representan un reto los tipos de muestras en la que los compuestos pertenecen a familias muy diferentes. En tales casos, no siempre es factible encontrar modelos matemáticos que ajusten la muestra. No es así, por ejemplo, cuando se trabaja en series homólogas, en las que casi cualquier método matemático

Tabla 3. Características de las nuevas muestras para clasificar los compuestos en activos e inactivos.

Muestra	k	AlgPreProc	Número de casos
FSM1	1	SMOTE	2 514
FSM5	5	SMOTE	2 514
FSM1CNN	1	SMOTE-CNN	651
FSM5CNN	5	SMOTE-CNN	846

Muestra; Nombre de la muestra, k; Valor del índice k empleado por el algoritmo SMOTE; AlgPreProc; Algoritmo de pre procesamiento; Número de Casos; Número total de casos después de los procesamientos correspondientes.

Tabla 2. Resumen de los datos correspondientes a los fragmentos empleados.

Fragmento	Valor mínimo	Valor máximo
Camino de orden 2	1,559 9	33,600 8
Camino de orden 3	- 0,395 7	35,048 9
Camino de orden 4	0,0	39,015 2
Ciclo de orden 3	0,0	20,026 2
Ciclo de orden 4	0,0	19,469 4
Ciclo de orden 5	0,0	33,789 6
Ciclo de orden 6	0,0	52,447 3
Ciclo de orden 7	0,0	39,136 3
Ciclo de orden 8	0,0	40,372 7
Ciclo de orden 9	0,0	40,798 1
Cluster de orden 3	- 5,352 3	45,353 6
Cluster de orden 4	- 1,435 7	41,311 3
Combinación cluster 3 - camino 2	0,0	68,858 9
Combinación cluster 3 - camino 3	0,0	85,037 5
Combinación cluster 4 - camino 2	0,0	56,110 5
Combinación cluster 4 - camino 3	0,0	65,146 2

Los valores de cero reportados indican la ausencia del fragmento en al menos una molécula del conjunto.

Tabla 4. Resultados de las predicciones calculadas mediante los diferentes algoritmos.

Muestra	AdaBoost	LogiBoost	MaxLogitBoost	Chi-RW	SGERD	Media
FSM1	0,67	***0,71***	0,59	0,68	0,54	0,64
FSM5	0,68	***0,71***	0,64	0,69	0,54	0,65
FSM1CNN	0,62	0,58	0,56	0,59	0,55	0,58
FSM5CNN	0,62	0,61	0,57	0,62	0,53	0,59
Media	0,65	0,65	0,59	0,64	0,540	

resulta válido para el establecimiento de modelos de relación entre la estructura química y la actividad biológica o la propiedad químico física que se evalúa. Hoy día es fácil encontrar múltiples ejemplos de estimación de propiedades en familias de sustancias congenéricas en las que se emplean métodos de regresión, redes neuronales, etc., con coeficientes de correlación por encima de 0,90 y bajos valores de error. No obstante, excepto para MaxLogitBoost y SGERD, los restantes algoritmos presentan valores promedio similares y se alcanza la mejor clasificación utilizando el algoritmo *LogiBoost* para cualesquiera de los valores de k empleados por el método SMOTE. Si se considera el *AdaBoost* como un modelo aditivo generalizado, al aplicar como función de costo una regresión logística, se tiene el algoritmo *LogiBoost*, que puede verse entonces como una optimización convexa en la que el algoritmo minimiza la pérdida de la función de costo logística.

Comparación entre algoritmos

Para poder establecer una diferencia clara entre los resultados, se aplicaron diferentes técnicas estadísticas. Inicialmente, se empleó la prueba de Friedman que asigna en este caso, el orden mayor al mejor de los algoritmos y viceversa (Tabla 5). Esta asignación se efectuó bajo el criterio de la hipótesis nula, la cual se forma a partir de suponer que los resultados de los algoritmos son equivalentes y, por tanto, sus *rankings* son similares.

Esta prueba estadística confirmó parcialmente el supuesto planteado en el apartado anterior acerca de que los mejores resultados se obtienen mediante los algoritmos Chi-RW, LogitBoost y AdaBoost. No obstante, la prueba de Friedman resultó significativa (0,01) y por lo tanto, se puede afirmar que existen diferencias globales significativas entre los algoritmos, entre los cuales se destacan el Chi-RW y LogiBoost como los de mejores resultados según el lugar que ocupan en el *ranking*. Sin embargo, no se distinguieron diferencias significativas entre esos tres algoritmos después de realizar una prueba de Wilcoxon de comparación por parejas para detectar diferencias más precisas mediante comparaciones de algoritmos dos a dos entre LogiBoost, AdaBoost y Chi-RW (Tabla 6). Si la hipótesis nula es cierta, los valores de los estadígrafos R+ y R- deberán ser parecidos, mientras que, a partir de nuestros datos, al ser más altos que la mediana M0 se refleja en un valor mayor de R+ y a la inversa si son más bajos. Esto se confirma con los valores de p que mide si los valores se distribuyen simétricamente alrededor del punto central según el nivel de significación elegido.[55]

Se pudo observar que, los tres algoritmos presentaban resultados equivalentes desde el punto de vista estadístico y que por lo tanto, pueden emplearse indistintamente.

Análisis de las diferentes muestras

Los resultados sugirieron que los mejores resultados se logran cuando se emplea $k = 5$ (Tabla 4), lo cual se pudo corroborar al graficar las medias de la predicción para los diferentes algoritmos y sus valores correspondientes de error y desviación estándar (Fig. 1).

Este resultado lo confirmaron además, los de las pruebas de Friedman y de Wilcoxon (Tablas 7 y 8), los cuales mostraron que las diferencias fueron significativas para un nivel de confianza de 0,05 y que entre las muestras FSM1 y FSM5, esta última presenta los mejores valores de predicción.

Tabla 5. *Rankings* obtenidos mediante la prueba de Friedman.

Algoritmo	*Ranking*
Chi-RW	4,25
LogiBoost	4,00
AdaBoost	3,75
MaxLogiBoost	2,00
SGERD	1,00
Significación	0,01

Tabla 6. Resultados de la evaluación de las predicciones con la prueba de Wilcoxon.

Comparación	R+	R-	*p*	Hipótesis ($\alpha = 0,05$)
LogiBoost *vs.* AdaBoost	5,00	5,00	1,000	Acepta H_0
LogiBoost *vs.* Chi-RW	7,00	3,00	0,715	Acepta H_0
AdaBoost *vs.* Chi-RW	4,00	6,00	0,465	Acepta H_0

R+; suma de los rangos de diferencia positiva; R-; suma de los rangos de diferencia negativa.

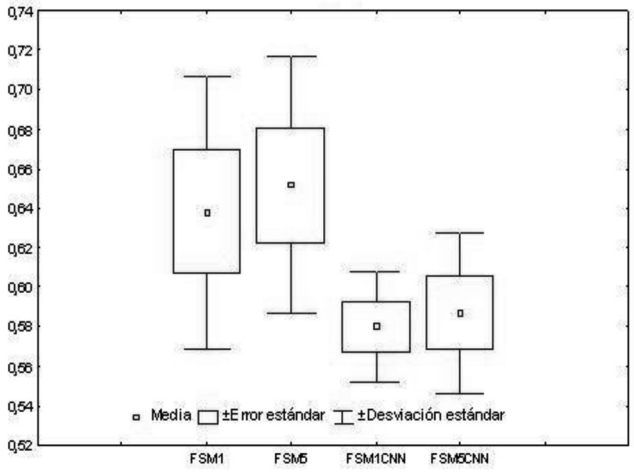


Fig. 1. *Comparación de medias, error estándar y desviación estándar para la predicción por fragmentos ponderados.*

Los resultados alcanzados permiten afirmar, que el mejor modelo se obtiene con el empleo de los algoritmos Chi-RW, AdaBoost y LogiBoost. Se aprecia también que es posible desarrollar modelos de clasificación en una muestra altamente desproporcionada en las clases con el empleo del algoritmo de pre procesamiento SMOTE con un valor de $k = 5$ aunque las diferencias con otros algoritmos no son estadísticamente significativas y la utilización de fragmentos ponderados por el Índice del Estado Refractotopológico Total para Fragmentos como variables independientes. Esta forma de descripción de la estructura química mostró resultados satisfactorios para el desarrollo de modelos de relación estructura -actividad. Con el empleo de los diferentes fragmentos moleculares ponderados por $\Re_T$ es posible lograr un mejor acercamiento a los detalles estructurales que inciden en la respuesta biológica, y por ende, facilitar la comprensión de la fenomenología analizada.

CONCLUSIONES

Se desarrollaron modelos de clasificación basados en algoritmos difusos evolutivos, mediante el empleo de un conjunto de estructuras moleculares con una gran diversidad estructural y un elevado índice de desbalance entre las clases. Los mejores modelos cualitativos se obtienen al balancear la muestra por el procedimiento SMOTE con una capacidad de predicción del 70 % para el algoritmo LogiBoost, el cual resultó estadísticamente equivalente a Chi-RW y AdaBoost. Estos muestran una calidad de clasificación media de 65,6, 65,3 y 64,8 % respectivamente, lo cual se considera un resultado aceptable dada la heterogeneidad de la muestra.

Se definió el Índice del Estado Refractotopológico Total para Fragmentos $\Re_T$, el cual es capaz de diferenciar fragmentos topológicamente iguales a partir del valor del índice, tanto intra como intermoleculares.

Tabla 7. Valores de *ranking* obtenidos mediante la prueba de Friedman.

Muestra	*Ranking*
FSM5	3,80
FSM1	2,80
FSM1CNN	1,80
FSM5CNN	1,60
Significación	0,026

Tabla 8. Resultados de la evaluación de las predicciones mediante la prueba de Wilcoxon.

Comparación	R+	R-	*p*	Hipótesis ($\alpha = 0,05$)
FSM5 *vs.* FSM1	15,00	0,00	0,043	Rechaza H_0 a favor de FSM5

R+; suma de los rangos de diferencia positiva; R-; suma de los rangos de diferencia negativa.

REFERENCIAS BIBLIOGRÁFICAS

1. Carrasco-Velar, R. Nuevos Descriptores Atómicos y Moleculares para Estudios de Estructura-Actividad. Aplicaciones Tesis para optar por el grado de Doctor en Ciencias Químicas. La Habana 2003.
2. Benson SW, Buss JH. J Chem Phys. 1958;29:546-58.
3. Hammett LP. The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. J Am Chem Soc. 1937;(59):96-103.
4. Kier LB, Hall LH. Molecular Connectivity in Chemistry and Drug Research. New York: Academic Press 1976.
5. Kier LB, Hall LH. Molecular Connectivity in Structure-Activity Analysis. London: John Wiley 1986.
6. Trinajstic N. Chemical Graph Theory. Boca Ratón, Fl: CRC 1983.
7. Estrada E. Three-Dimensional Molecular Descriptors Based on Electron Charge Density Weighted Graphs. J Chem Inf Comput Sci. 1995(35):708-13.
8. Estrada E. Edge Adjacency Relationships and a Novel Topological Index Related to Molecular Volume. J Chem Inf Comput Sci. 1995(35):31-3.
9. Estrada E, LA. Montero. Bond Order Weighted Graphs in Molecules as Structure-Property Indices. Moleng. 1993(2):363-73.
10. Estrada E, Molina E. 3D Connectivity Indices in QSPR/QSAR Studies. J Chem Inf Comput Sci. 2001(41):791-7.
11. Estrada E, Molina E, Perdomo I. Can 3D Structural Parameters be Predicted from 2D (Topological) Molecular Descriptors? J Chem Inf Comput Sci. 2001(41):1015-21.
12. Carrasco-Velar R, Padrón JA, Gálvez J. Definition of a Novel Atomic Index for QSAR: The Refractotopological State. J Pharm Pharmaceut Sci. (www.ualberta.ca/~Csps). 2004;7(1):19-26.
13. Padrón-Garcia JA, Carrasco-Velar R, Pellón R. Molecular Descriptor Based on a Molar Refractivity Partition Using Randic-Type Graph-Theoretical Invariant. J Pharm Pharmaceut Sci. (www.ualberta.ca/~Csps). 2002;5(3):267-74.

14. Ertl P, S. Sj. Curr Top Med Chem. 2007(7):1491-516.
15. Kühne R, Ralf-Uwe E, Schuüürmann G. Chemical Domain Of QSAR Models From Atom-Centered Fragments. J Chem Inf Model. 2009(49):2660-9.
16. Zadeh LA. Fuzzy Logic, Neural Networks and Soft Computing. Communications of the ACM. 1994:77-8.
17. Piñero PY. Un Modelo para el Aprendizaje y la Clasificación Automática Basado en Técnicas de Softcomputing Tesis para optar por el grado de Doctor en Ciencias Técnicas Universidad Central de Las Villas; 2005.
18. Herrera F. Sistemas Difusos Evolutivos. Jaen, Universidad de Jaén; 2004.
19. Borgelt C, Kruse R. Learning Possibilistic Graphical Models from Data. Fuzzy Systems. IEEE Transactions. 2003;11(2):159-72.
20. Rutkowski L. Flexible Neuro-Fuzzy Systems. Structures, Learning and Performance Evaluation: Kluwer Academic Publishers 2004.
21. Herrera F, Alcalá R. Genetic Tuning on Fuzzy Systems Based on the Linguistic 2-Tuples Representation. IEEE International Conference on Fuzzy Systems (Fuzz-IEEE04), Budapest, Hungary. 2004:233-8.
22. Ishibuchi H, Nakashima T, Murata T. Performance Evaluation of Fuzzy Classifier Systems For Multidimensional Pattern Classification Problems. IEEE Transactions on Systems and Man and Cybernetics and Part B: Cybernetics. 1999(29):601-18.
23. Pedrycza W, Reformata M. Genetically Optimized Logic Models. Fuzzy Sets and Systems. 2005(150):351-71.
24 Cordón O, Gomide F, Herrera F, Hoffmann F, Magdalena L. Ten Years of Genetic Fuzzy Systems. Current Framework and New Trends. Fuzzy Sets and Systems. 2004(41):5-31.
25. Batista G, Prati R, Monard MC. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. ACM Sigkdd Explorations. 2004;6(1):20-9.
26. Moreno J, Rodríguez D, Sicilia MA, Riquelme JC, Ruiz R. Actas de los Talleres de las Jornadas de Ingeniería del Software y Bases de Datos, SISTEDES. 2009;3(1):73.
27. Tomek I. An Experiment with the Edited Nearest-Neighbor Rule IEEE Transactions On Systems, Man and Cybernetics. 1976(SMC-6):448-52.
28. Tomek I. Two Modifications of CNN. IEEE Transactions on Systems Man and Communications. 1976(SMC-6):769-72.
29. Fuzzy Systems and Knowledge Discovery. *Lecture Notes In Computer Science.* 2005:39-48.
30. Morales EF, González JA. El Problema de las Clases Desbalanceadas. Aprendizaje Computacional II. 2009 Cited: 2010; 3 de Mayo; Escuela Verano, Material De Estudio. Instituto Nacional de Astrofísica, Óptica y Electrónica, . Available From: Http://ccc.inaoep.mx/~jagonzalez/ml2/ml2.html
31. Bradley P. The Use of the Area Under the Roc Curve In the Evaluation of Machine Learning Algorithms,. Journal Pattern Recognition. 1997(30):7.
32. Riquelme JC, Ruiz R, Rodríguez D, Moreno J. Actas de los Talleres de las Jornadas de Ingeniería del Software y Bases de Datos, SISTEDES. 2008;2(1).
33. Hart P. The Condensed Nearest Neighbour Rule, IEEE Transactions on Information Theory. 1968(14):515-6.
34. Kubat M, Matwin S. Addressing the Curse of Imbalanced Training Sets: One Sided Selection. Proc ICML-97, Nashville (USA). 1997:179-86.
35. Chawla NV, Bowyer K, Hall L, Kegelmeyer W. SMOTE: Synthetic Minority Over-Sampling Technique. Artificial Intelligence Research. 2002(16):321-57.
36. Burez J, Poel D. Handling Class Imbalance in Customer Churn Prediction. Expert Systems with Applications. 2009(36):4626-36.
37. Fernández J, Herrera F. Hierarchical Fuzzy Rule Based Classification Systems with Genetic Rule Selection for Imbalanced Data-Sets. International Journal of Approximate Reasoning. 2009(50):561-77.
38. Fernández-García S, Jesús MD, Herrera F. A Study of the Behaviour of Linguistic Fuzzy Rule Based Classification Systems in the Framework of Imbalanced Data-Sets. Fuzzy Sets and Systems. 2008(159):2378-98.
39. Hall LH, Mohney B, Kier LB. The Electrotopological State: Structure Information at the Atomic Level for Molecular Graphs. J Chem Inf Comput Sci. 1991;31(1):76-82.
40. Ghose AK, Crippen GM. Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships. I. Partition Coefficients as a Measure of Hydrophobicity. J Comput Chem 1986;7(4):565-77.
41. Ghose AK, Crippen GM. Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships. 2. Modeling Dispersive and Hydrophobic Interactions. J Comput Chem. 1987;27(1):21-35.
42. Jha T, Samanta S, Basu S, Kumar-Halder A, Adhikari N, Kumar-Maiti M. QSAR Study on Some Orally Active Uracil Derivatives as Human Gonadotropin-Releasing-Hormone Receptor Antagonists. Internet Electron J Mol Des. 2008;7(11):234-50.
43. Samanta S, Alam SM, Panda P, Jha T. Pharmacophore Mapping of Tricyclic Isoxazoles for their Affinity Towards Alpha-2 Adrenoreceptors. *Internet Electron J Mol Des* 2006:503-14.
44. Último acceso:12/11/2010 Http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=435&loc=ea_ras
45. Orriols-Puig E, Bernadó-Mansilla K, Sastry J, Goldberg De. Substructural Surrogates for Learning Decomposable Classification Problems: Implementation and first Results. New York: Acm Press, 2007.
46. Alcalá-Fernández J, Sánchez L, García Md, Jesús S, Ventura J, Garrell J, *et al*. Keel: A Software Tool to Assess Evolutionary Algorithms for Data Mining Problems Soft Computing. 2009(13):307-18.
47. Chi Z, Yan H, Pham T. Fuzzy Algorithms: With Applications to Image Processing and Pattern Recognition. Singapore: World Scientific Publishing Co. Pte. Ltd. 1996.
48. Mansoori E, Zolghadri M, Katebi S. SGERD: A Steady-State Genetic Algorithm for Extracting Fuzzy Classification Rules from Data,. IEEE Transactions on Fuzzy Systems. 2008;16 (4):1061-71.
49. M.J. Del Jesús, Hoffmann F, Junco L, Sánchez L. Induction of Fuzzy-Rule-Based Classifiers with Evolutionary Boosting Algorithms. IEEE Transactions on Fuzzy Systems. 2004(12:):296-308.
50. Otero J, Sánchez L. Induction of Descriptive Fuzzy Classifiers with the Logitboost Algorithm. Soft Computing. 2006(10):825-35.
51. Sánchez L, Otero J. Boosting Fuzzy Rules in Classification Problems Under Single-Winner Inference. International Journal of Intelligent Systems. 2007(22):1021-34.
52. SPSS. Statistical Package For The Social Sciences. 13.1 Ed.
53. Stewart JJP. Optimization of Parameters for Semiempirical Methods IV: Extension of Mndo, AM1, and PM3 to More Main Group Elements. J Mol Mod. 2004;2(10):155-64.
54. Universidad de La Habana Cuba. Mopac, V. 6.0, Release 1.02.1997:for 3/486/Pentium PC'S. Windows 95 and NT Enviroments.
55. Molinero LM. ¿Y Si Los Datos No Siguen Una Distribución Normal?. Bondad de Ajuste a una Normal. Transformaciones Pruebas no Paramétricas 2003 accesado: 2010 19/8; Http://www.seh-lelha.org/noparame.htm