Asociación Argentina
de Mecánica Computacional

Mecánica Computacional Vol XXVII, págs. 2429-2441 (artículo completo)
Alberto Cardona, Mario Storti, Carlos Zuppa. (Eds.)
San Luis, Argentina, 10-13 Noviembre 2008

APLICACIONES DE APRENDIZAJE AUTOMÁTICO SOBRE CLUSTERS DE COMPUESTOS QUÍMICOS

Axel J. Soto[a,b], Damián Palomba[a], Mónica F. Diaz[b], Gustavo E. Vazquez[a] e Ignacio Ponzoni[a,b]

[a]*Laboratorio de Investigación y Desarrollo en Computación Científica (LIDeCC), Departamento de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur, Alem 1253, 8000 Bahía Blanca, Argentina, ip@cs.uns.edu.ar, http://www.lidecc.cs.uns.edu.ar*
[b]*Planta Piloto de Ingeniería Química (PLAPIQUI), UNS-CONICET, Complejo CCT, Camino La Carrindanga km. 7, 8000 Bahía Blanca, Argentina, asoto@plapiqui.edu.ar, http://www.plapiqui.edu.ar*

Palabras clave: Quimioinformática, QSAR, Aprendizaje Supervisado, Aprendizaje No Supervisado.

Resumen. Las técnicas de inteligencia computacional o aprendizaje automático (*machine learning*) son de actual relevancia para el análisis y estudio de propiedades en compuestos químicos. En particular, para aquellos compuestos que son medicamentos, drogas o candidatos a drogas, resulta de vital importancia contar con técnicas computarizadas que asistan en la predicción de propiedades farmacocinéticas.

Una de las técnicas de vigente estudio y aplicación en el área de quimioinformática es QSAR/QSPR (*Quantitative Structure Activity/Property Relationships*). La misma consiste en el análisis de las relaciones existentes entre la estructura molecular de un compuesto químico y una determinada actividad o propiedad biológica. En este tipo de técnicas un compuesto químico tiene asociado un número de descriptores moleculares, en donde cada descriptor representa una determinada característica del compuesto.

Por otra parte, dado que el espacio de posibles compuestos químicos capaces de ser sintetizados es extremadamente grande, las técnicas de análisis de agrupamientos (*cluster analysis*) resultan ser una herramienta de interés para la mejora en el entendimiento de las relaciones entre estructura y propiedad.

El presente trabajo tiene por objetivo analizar y mostrar las ventajas en la incorporación de clustering como parte del proceso de predicción de la hidrofobicidad de un compuesto. Se realizaron distintas alternativas de agrupamiento, procurando que dicha asociación sea de relevancia para la propiedad a modelar.

Para la tarea de predicción numérica de la hidrofobicidad, las redes neuronales es la herramienta mayormente usada. Los resultados obtenidos ponen en evidencia las ventajas de la división del conjunto muestral de datos en subconjuntos de menor tamaño, y la utilización de esa división en la tarea de predicción.

1. INTRODUCCIÓN

Históricamente, el desarrollo de nuevos productos farmacéuticos consistía en un proceso de prueba y error que se basaba mayormente en la búsqueda de un farmacóforo o principio activo. Sin embargo, muchos compuestos eran finalmente descartados debido al comportamiento ADMET (*Absorption, Distribution, Metabolism, Excretion and Toxicity*) dentro del cuerpo humano (Selick et al., 2002). El estudio de las propiedades farmacocinéticas de las drogas en un estadío temprano del desarrollo de drogas, y en paralelo con la búsqueda de una dada actividad, corresponde a un paradigma más moderno del proceso de fabricación de drogas. Este paradigma, permite disminuir los tiempos de desarrollo y por consiguiente el dinero invertido.

Por estos motivos, el interés dentro de la industria y el ámbito académico en las disciplinas de quimioinformática (*chemoinformatics*) y quimiometría (*chemometrics*) y en particular las técnicas de tipo QSAR (*Quantitative Structure Activity Relationships*), ha crecido considerablemente en los últimos años. Esta técnica se basa en la búsqueda de relaciones entre la estructura química de un compuesto y un determinado proceso, tal como una actividad biológica o una propiedad fisicoquímica. La información sobre la estructura química de un compuesto es representada por medio del cálculo de descriptores moleculares (Todeschini and Consonni, 2000; Livingstone, 2000), los cuales brindan valores cuantitativos sobre determinados aspectos de la molécula. Cuando estos métodos de predicción de propiedades, se realizan por medios computacionales, se denominan técnicas *in silico*.

Si bien estos métodos *in silico* no pretenden reemplazar a los métodos experimentales, al menos en el corto término, algunos métodos computacionales han demostrado tener tan buena precisión como métodos experimentales bien asentados (Agatonovic-Kustrin and Beresford, 2000). La principal ventaja de estos métodos es que permiten el análisis de una droga (o de toda una librería) sin necesidad de sintetizar la molécula.

Independientemente de los descriptores con los que se trabaje, cuando se consideran compuestos heterogéneos, el espacio químico, determinado por los valores de los descriptores moleculares con los que se define un compuesto, suele poseer un alta dimensionalidad y una gran dispersión. Por lo tanto, un inconveniente mayor que se presenta al momento de desarrollar un método basado en QSAR, surge de la gran diversidad de compuestos y la consecuente inmensidad del espacio químico asociado. Por consiguiente, no resulta trivial desarrollar modelos de predicción que sean lo suficientemente generales como para ser aplicados a cualquier compuesto químico (Tetko et al., 2006).

Por esta razón, el presente trabajo busca analizar y mostrar ventajas obtenidas en la incorporación del concepto de similaridad en el proceso de predicción de propiedades de un compuesto. En particular, analizaremos esta propuesta con la hidrofobicidad como propiedad fisicoquímica a predecir. Este trabajo se encuentra organizado como sigue: en la siguiente sección se describen algunos aspectos concernientes a la predicción de propiedades y el estado del arte del análisis de similaridad en la predicción, en la sección 3 se analizan distintos criterios de agrupamientos; la cuarta sección detalla un procedimiento propuesto por los autores que permite mejorar la generalización en la predicción de compuestos químicos y finalmente, en la sección 5 se resaltan los principales aportes y conclusiones.

2. PREDICCIÓN DE PROPIEDADES

Los métodos de predicción de propiedades *in silico* resultan más útiles cuanto más necesaria y difícil sea obtener experimentalmente dicha propiedad. Una de las propiedades fisicoquímicas mayormente modeladas es la hidrofobicidad, la cual refleja la afinidad de una molécula a

unirse al agua o a los lípidos. La justificación del interés que despierta esta propiedad, reside en que ésta influencia considerablemente el comportamiento de un xenobiótico dentro del cuerpo humano. Esta propiedad se la expresa en términos del logaritmo del coeficiente de partición octanol-agua (logP). A mayor valor de logP, la droga es más hidrofóbica. El valor de logP de un compuesto puede ser usado como un medio de rechazo de moléculas candidatas a drogas, en una etapa temprana del proceso de desarrollo de fármacos (Taskinen and Yliruusi, 2003).

Acorde a este interés en el modelado de la hidrofobicidad, y comenzando con el trabajo de Hansch and Leo (1979), muchos métodos basados en el enfoque QSAR han sido desarrollados en las últimas décadas. Se puede hacer una diferenciación entre los métodos basados en contribuciones grupales y los basados en el cálculo de descriptores moleculares (Erös et al., 2002). Los primeros plantean una estimación del valor de la propiedad a partir del aporte individual de determinados átomos o fragmentos. Los segundos se cimentan en la construcción de un modelo parametrizado por los valores de los descriptores considerados. Un descriptor es el resultado de algún experimento o procedimiento matemático para extraer información de un compuesto químico *e.g.* peso molecular, superficie de la molécula, cantidad de carga electroestática, etc. Por consiguiente, aquellos métodos basados en descriptores constituyen una alternativa más interesante y general.

Dentro de las ciencias de la computación, una disciplina en auge en estos días es el aprendizaje automático (*machine learning*), la cual consiste en el reconocimiento de reglas y patrones que surgen del análisis de grandes cantidades de datos. El aprendizaje automático, básicamente, se divide en dos grandes modalidades: supervisado y no supervisado. En el caso del aprendizaje supervisado, lo que se plantea es modelar un sistema en base a datos de entrada y de salida del mismo. Se debe disponer de suficiente cantidad de información para que el modelo a construir permita aprender sobre la información provista, y así representar el comportamiento deseado.

Por otra parte, en el aprendizaje no supervisado el objetivo es capturar características relevantes de los datos presentados. Cabe mencionar que en esta modalidad no existe una distinción entre datos de entrada y de salida de un sistema, ya que en este caso no se busca aprender su comportamiento. En la siguiente sección, expandiremos este tipo de aprendizaje mediante la introducción de un método particular.

En la actualidad, los métodos de aprendizaje automático aplicados a la predicción de propiedades, resultan de destacada importancia. La combinación de aprendizaje supervisado e información suficiente de compuestos químicos, permite obtener un modelo que captura en forma automatizada las relaciones estructura-propiedad. En este contexto, las redes neuronales (*neural networks - NN*) o los comités de redes neuronales (*neural network ensembles - NNE*) son en la actualidad uno de los métodos de predicción más apropiados para este tipo de aprendizaje (Taskinen and Yliruusi, 2003; Winkler, 2004).

2.1. Dificultades en los métodos de predicción de logP

El desafío de un método de predicción reside en la capacidad de obtener modelos lo suficientemente generales como para que el mismo modelo funcione sobre compuestos no presentados anteriormente. Un causante de esta dificultad es la incertidumbre sobre la cantidad y la elección del conjunto de descriptores necesarios para predecir una propiedad. Excluyendo a unos pocos descriptores, no existe un consenso general sobre cuáles son los descriptores que influyen en la predicción de la hidrofobicidad.

Asimismo, la generalización es un problema mayor al usarse datos no homogéneos *i.e.* el poseer grupos de datos sobre-representados provocará que se modele en demasía estos elementos, en detrimento de los que no pertenezcan al grupo popular (concepto estadístico conocido como

sobreajuste). Jónsdottir et al. (2005) mencionan este problema y la necesidad de que sea tenido en cuenta en la construcción de los modelos.

Al mismo tiempo, pareciera que aún se está lejos de la obtención del predictor de logP universal, es decir, un método confiable que permita predecir logP para cualquier compuesto presentado (Tetko et al., 2006). Prueba de ésto es que modelos promisorios fallaron al ser testeados con compuestos externos a los usados en el entrenamiento (Tetko, 2002b).

Numerosas propuestas utilizan el concepto de similaridad de compuestos químicos. En Willet (2000) se detalla el interés por identificar diversidad en compuestos químicos de una manera eficiente, para de este modo mejorar la búsqueda de nuevas formas farmacéuticas. En Espinosa et al. (2002) el objetivo del análisis de similaridad, pasa por identificar el conjunto de descriptores óptimos para desarrollar un modelo de predicción. Otros trabajos (Kühne et al., 2006; Sheridan et al., 2004; Tetko, 2002a) atacan los problemas de la redundancia y la extrapolación con medidas de similaridad, para desarrollar modelos más robustos. En tal sentido, nuestras investigaciones también apuntan a la utilización de conceptos de agrupamiento para realizar un estudio exploratorio de la aplicabilidad de un modelo y así servir de base para una mejora de la capacidad predictiva del mismo.

3. ANÁLISIS DE AGRUPAMIENTO

El análisis de agrupamientos (*cluster analysis*) consiste en el particionamiento de un conjunto de elementos en grupos (*clusters*) en donde se busca que los elementos pertenecientes a un mismo grupo guarden alguna medida de similitud y, al mismo tiempo, mantengan diferencias con los elementos conformados en otros grupos. Esta búsqueda de características similares dentro de un conjunto de datos, provee una herramienta de interés al momento de realizar un análisis exploratorio, como es la búsqueda de relaciones entre los datos o la posibilidad de asignación de un elemento como representativo de un grupo.

Un aspecto crucial en el análisis de agrupamiento es definir con qué criterio es calculada la similitud o diferenciación entre dos elementos o grupos. En la siguiente sección detallaremos los enfoques de agrupamiento con los que se experimentó en el presente trabajo.

3.1. Usando métricas de distancias

Existen diferentes métodos de análisis de agrupamiento basados en el establecimiento de medidas de distancias. Considerando a cada compuesto representado en un espacio $\mathbb{R}_n$, definido por el valor de los n descriptores moleculares, se puede medir su similaridad mediante el cálculo de la distancia entre un punto y otro. La distancia entre individuos puede ser medida usando una distancia Euclídea; o bien otros tipos de distancias *e.g.* Gower, Canberra, cuerda de Orloci o distancia coseno. En el caso de descriptores binarios, son otras las medidas de distancia que corresponden ser usadas (Dice, Tanimoto, *simple matching*). La elección por una distancia u otra será en función del tipo de datos y del criterio con que se pretende diferenciar a los elementos.

Dentro de esta clase de técnicas de agrupamiento existen, básicamente, dos tipos de métodos: jerárquicos y de partición (Johnson and Wichern, 1992). En el caso de los primeros, se realizan sucesivos ligamientos (enfoque aglomerativo) o separaciones (enfoque divisivo). De esta forma, no existe una única manera de congregar elementos ya que el agrupamiento puede realizarse a distintos niveles de similitud.

Los métodos de partición, fueron pensados para agrupar elementos alrededor de k puntos centrales, donde k es un parámetro especificado *a priori*. Cada elemento se une al grupo, que mantenga una menor distancia con el centro. Estos métodos son de provecho para trabajar con

grandes cantidades de datos, ya que no requieren el almacenamiento de la distancia entre cada par de elementos. Sin embargo, requieren de la definición de puntos centrales que sean apropiados para el problema en cuestión.

3.2. Usando aprendizaje automático

En la sección anterior, detallamos las diferencias entre los modos (supervisado y no supervisado) de aprendizaje automático. El aprendizaje no supervisado comúnmente se utiliza para identificar grupos de datos en forma automática. Esta identificación puede realizarse a nivel de las variables o a nivel de las elementos. En el primer caso, el objetivo es encontrar un subconjunto de variables que sean representativas y descartar aquellas que sean redundantes. En la identificación a nivel de los elementos, el objetivo es el mismo al planteado cuando se introdujo el análisis de agrupamiento.

Dentro de las técnicas más comunes de aprendizaje no supervisado encontramos a los mapas auto-organizativos (*self organizing maps - SOMs*) y a las LVQ (*learning vector quantization*) (Kohonen, 1997). Ambas presentan un comportamiento análogo, pero nos centraremos en los SOMs, dado que es el método de agrupamiento usado en este trabajo.

El SOM es un algoritmo generalmente agrupado dentro de las redes neuronales. Los nodos o celdas de salida suelen estar dispuestos en formas geométricas, donde cada nodo actúa de modo competitivo ante la presentación de un caso. A su vez, cada nodo posee un vector de pesos, de igual dimensión que los datos de entrada. La figura 1 muestra un esquema de la arquitectura de un SOM con los nodos dispuestos en forma de grilla rectangular de 9×7 y donde los casos presentados poseen n dimensiones.

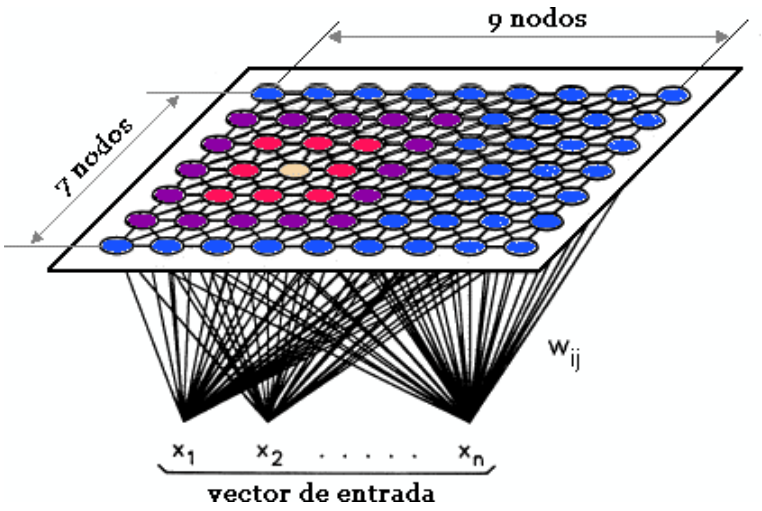

Figura 1: Arquitectura de un SOM.

Cada caso es presentado al SOM, en forma aleatoria y consecutiva, en donde el nodo ganador (NG) resulta de comparar el caso presentado con los vectores de pesos de cada nodo de la grilla, definiéndose al ganador siguiendo alguna medida de distancia. El peso del NG es modificado de manera de tal que se disminuya la distancia al caso presentado. La ecuación 1 refleja esta actualización de los pesos para el NG, en donde t representa el instante de tiempo y L es el

ritmo de aprendizaje, el cual decae con el tiempo.

$$W(t+1) = W(t) + L(t)(V(t) - W(t)), \qquad L(t) = L_0 e^{-\frac{t}{\lambda}}. \tag{1}$$

Una característica distintiva de los SOMs es la capacidad de preservación de la topología. Por tanto, cada nodo guarda una relación de vecindad con los nodos próximos de la grilla, de manera que al actualizarse el peso del NG, también se actualizan los pesos de los nodos vecinos. La forma de calcular la vecindad varía de un algoritmo a otro, pero generalmente el alcance de la vecindad se achica con el avance del algoritmo. La ecuación 2 muestra una posible formulación para establecer la vecindad en el tiempo.

$$\sigma(t) = \sigma_0 e^{-\frac{t}{\lambda}}. \tag{2}$$

Finalmente, cabe decir que el impacto en la actualización de los pesos de los nodos vecinos a un NG, varía en función de la distancia que se encuentre del mismo, *i.e.* dentro del área de vecindad, el cambio en el peso de un nodo es mayor cuanto más cerca se esté del NG. Con la incorporación del concepto de vecindad, la actualización de los pesos para cualquier nodo dentro de $\sigma(t)$, puede ser representado según la ecuación 3 en donde Θ representa el grado de influencia de la distancia del nodo al NG.

$$W(t+1) = W(t) + \Theta(t)L(t)(V(t) - W(t)), \qquad \Theta(t) = e^{-\frac{dist^2}{2\sigma^2(t)}} \tag{3}$$

3.3. Usando criterios definidos por expertos

Como un último enfoque, surgió el interrogante de si resulta viable establecer un criterio basado en el análisis humano. Para este trabajo, se dividió el conjunto de compuestos considerando las características químicas definidas por la presencia o ausencia de grupos funcionales. La división propuesta es similar a la planteada en Yaffe et al. (2002) para el análisis de la salida; con la salvedad de que en nuestro caso se fusionaron aquellos grupos que poseían baja cardinalidad.

En el primer grupo, se concentraron sustancias que sólo estaban constituidas por átomos de carbono e hidrógeno. Para el segundo y tercer grupo, se consideraron sustancias que contuvieran grupos carboxilo y grupos carbonilo respectivamente. Al cuarto grupo, se asignaron las moléculas que constaban de un grupo oxhidrilo unido a una cadena abierta o cíclica; o directamente unida a un átomo de carbono aromático (SP^2). En el grupo 5, aglomeramos a todos los compuestos con azufre junto a los éteres, debido a que ambos mantienen una similaridad estructural. El sexto grupo, se conformó por moléculas que contienen un átomo de nitrógeno unido a hasta tres grupos alquilo, o bien a un anillo aromático. Las sustancias que tienen al menos un átomo halógeno unido a un grupo alquilo o a un átomo de carbono aromático fueron asignadas al grupo 7. Finalmente, el último grupo se confeccionó con aquellas moléculas que contienen un anillo formado por más de un tipo de átomo (nitrógeno y oxígeno, además de carbono).

4. DETALLE DE LAS EXPERIMENTACIONES

El propósito del presente artículo, es utilizar enfoques de análisis de agrupamiento para el uso conjunto con métodos de predicción basados en QSAR. La motivación de nuestra hipótesis

de trabajo surge de expandir la idea presente en la literatura de quimioinformática que, la identificación de características similares dentro de un conjunto de compuestos químicos, provee información que puede ser usada para la mejora de las relaciones estructura propiedad (Martin et al., 2002). Al mismo tiempo, la vinculación de diferentes métodos de aprendizaje automático (Wolpert, 1992; Gama and Brazdil, 2000) en un mismo método, permite, generalmente, disminuir los errores asociados a un método de predicción. En trabajos anteriores (Soto et al., 2007) hemos desarrollado un método de predicción en base a un algoritmo de agrupamiento jerárquico. Por tanto, aquí sólo incluiremos resultados utilizando las técnicas desarrolladas en la Sección 3.2 y 3.3.

4.1. Separación por SOMs

Comenzaremos nuestro análisis experimentando con los SOMs. El mapa fue entrenado durante 500 generaciones, en donde el entrenamiento es dividido en dos fases. En la primera, la tasa de aprendizaje comienza con 0.9 (L_0 en Ecuación 1) hasta el comienzo de la segunda, en donde la tasa se fija en 0.02. Algo similar sucede con el alcance de la vecindad, donde en la primer fase σ_0 es igual a la máxima distancia entre dos nodos, y a partir de la segunda fase σ se mantiene en 1.

Los datos de los compuestos usados corresponden al trabajo de Yaffe et al. (2002), y los descriptores medidos para esos compuestos corresponden a los 34 mejores descriptores obtenidos en Soto et al. (2008), según un proceso de selección de descriptores sobre el mismo set de datos. La Figura 2 ilustra el resultado de entrenar las moléculas con un SOM, donde sus nodos están dispuestos en forma de una grilla cuadrada de 7×7. La distancia entre dos nodos se calcula mediante la distancia euclídea, donde las coordenadas de un nodo corresponden a su índice dentro de la matriz.

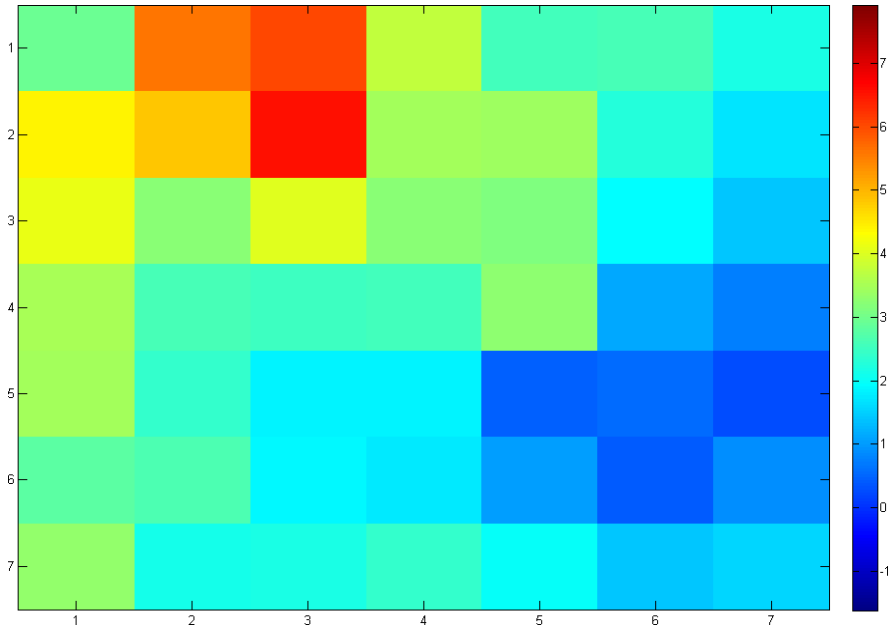

Figura 2: Promedio logP por NGs (7×7).

El color de un nodo, y su valor asociado en la barra de colores, proviene de aplicar el promedio de valor de logP de todos los compuestos que tienen a ese nodo como NG. Podemos ver en la Figura 2 que existe una homogeneidad de los valores de logP según las regiones del

mapa, *e.g.* en el borde inferior derecho vemos que se han agrupado compuestos hidrofílicos, mientras que en el borde superior izquierdo, lo han hecho los compuestos hidrofóbicos. Esto resulta interesante, teniendo en cuenta que el valor de logP de un compuesto no fue considerado al momento del entrenamiento del SOM.

Además de que el promedio de logP entre nodos vecinos mantenga una continuidad, resulta importante analizar, cuál es la homogeneidad dentro de cada nodo. Para esto, calculamos el desvío estándar dentro de cada celda (Figura 3). Según se aprecia, con excepción de algunos pocos nodos, dentro de cada nodo existe poca dispersión de valores de logP. No obstante, como se analizará más adelante, estos nodos con alta dispersión, pueden darnos información valiosa.

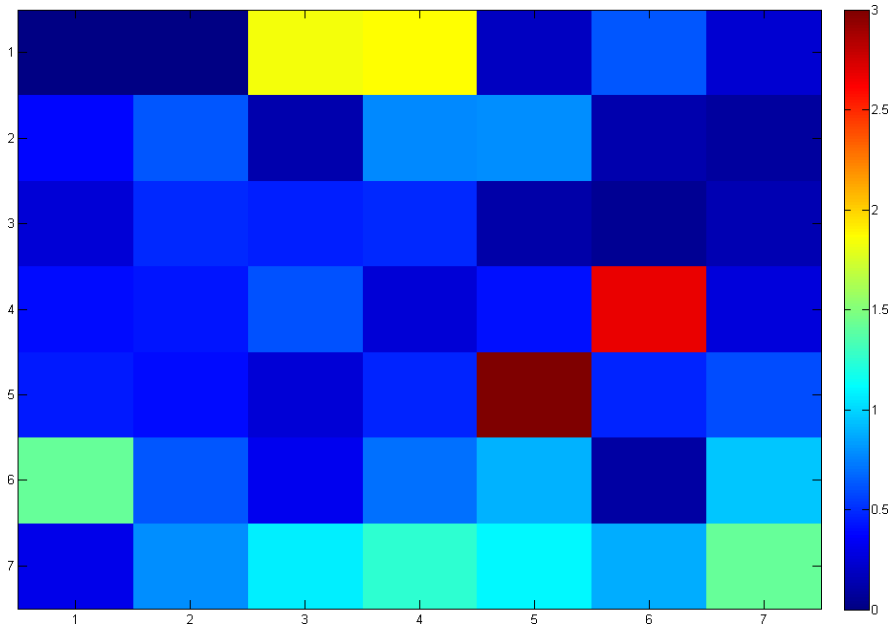

Figura 3: Desvío estándar logP por NGs (7×7)

El mismo análisis puede hacerse para mapas de distintas cantidades de nodos. Las Figuras 4 y 5 muestran los valores promedios de los compuestos entrenados con mapas de dimensión 10 y 15 respectivamente. Como se ve, la continuidad de valores entre nodos cercanos se mantiene. Es importante notar, que los mapas han tomado otra distribución en cuanto a la ubicación de los valores más altos y más bajos. Esto no invalida el método, y sucede debido a la aletoreidad existente tanto en los pesos iniciales como en el orden de elección de compuestos durante el entrenamiento del SOM. Asimismo, vemos que a medida que se aumenta el tamaño de la grilla aparecen más nodos que no quedan asignados a ningún compuesto (pintados en blanco).

Finalmente, complementamos el análisis anterior con la inclusión de las Figuras 6 y 7. Nuevamente se aprecia que, en general, la dispersión dentro de cada celda es baja.

4.2. Separación por clases químicas

En el caso de la separación por clases químicas, se produjo una distribución que fue menos homogénea respecto del caso anterior. La Tabla 1 refleja el análisis del promedio y la desviación asociada a cada grupo establecido. Como se ve, las clases químicas confeccionadas no guardan una relación con el grado de hidrofobia de los compuestos, ya que los promedios obtenidos son similares y el desvío estándar siempre es superior a $0{,}86$.

Por otra parte, se intentó también inferir relaciones de tipo estructura-actividad que pudieran

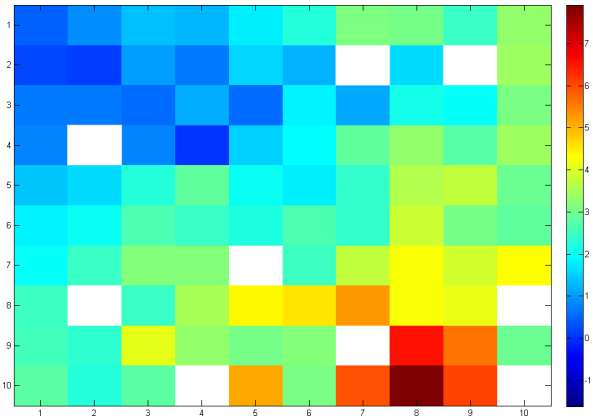

Figura 4: Promedio logP por NGs (10×10)

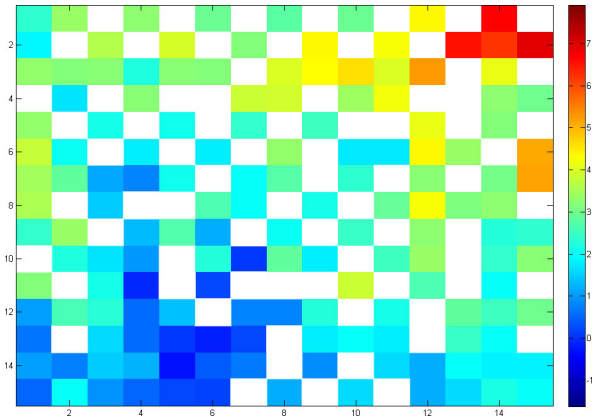

Figura 5: Promedio logP por NGs (15×15)

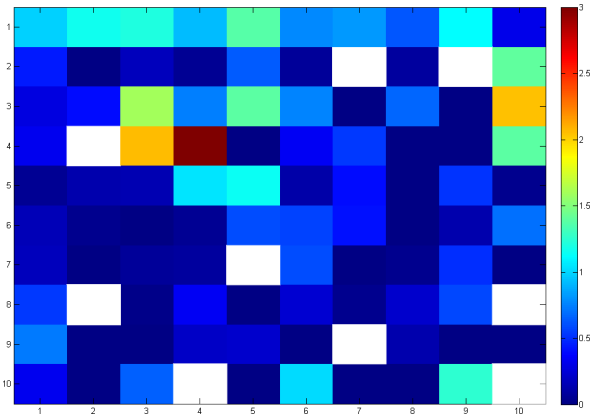

Figura 6: Desvío estándar logP por NGs (10×10)

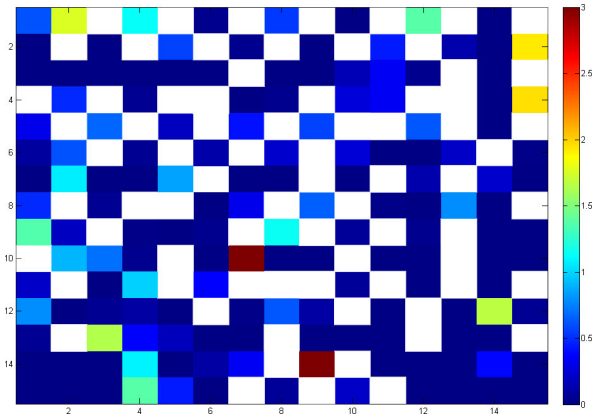

Figura 7: Desvío estándar logP por NGs (15×15)

desprenderse dentro de cada grupo. Esto no prosperó debido en parte a la cardinalidad desigual de cada grupo.

	Cantidad Comp.	Promedio logP	Desvío Est. logP
Grupo 1	117	3.2326	1.0690
Grupo 2	12	1.5658	1.4089
Grupo 3	32	1.4119	1.15
Grupo 4	57	1.8100	0.9575
Grupo 5	26	2.0392	1.2593
Grupo 6	37	1.4986	0.9921
Grupo 7	119	3.2050	1.6054
Grupo 8	24	1.3738	1.0199
Grupo 9	16	0.6556	0.8628

Tabla 1: Promedio y desviaciones por clases químicas.

4.3. Combinación de aprendizaje supervisado y no supervisado

A continuación detallaremos el desarrollo de nuestra hipótesis que plantea la aplicación de un método de agrupamiento combinado con un método de aprendizaje supervisado, en pos de mejorar el desempeño de las predicciones obtenidas. En particular se aplicó un modelo de NNE, en donde se utilizaron 3 NN sobre un 80 % de los datos seleccionados al azar. El 20 % restante

fue dejado para testeo (*hold-out validation*). Para estas experimentaciones se utilizaron redes de tipo *feed-forward* y el algoritmo de aprendizaje fue el Levenberg-Marquardt (Bishop, 2006). Utilizaremos también la red SOM de 49 nodos de la Sección 4.1.

En primer término, asignamos un valor a cada celda del SOM, como se hizo en la sección 4.1, pero con la diferencia que cada valor de los NGs será asignado de acuerdo al promedio de las predicciones del NNE de los compuestos pertenecientes a ese nodo (Figura 8). Podemos observar que visualmente, se obtiene una disposición semejante a la de la Figura 2. Esto nos indica, en un golpe de vista, que al menos los compuestos del entrenamiento, son modelados por la red correctamente. Sin embargo, es importante analizar cómo predice el NNE para los compuestos reservados para el testeo.

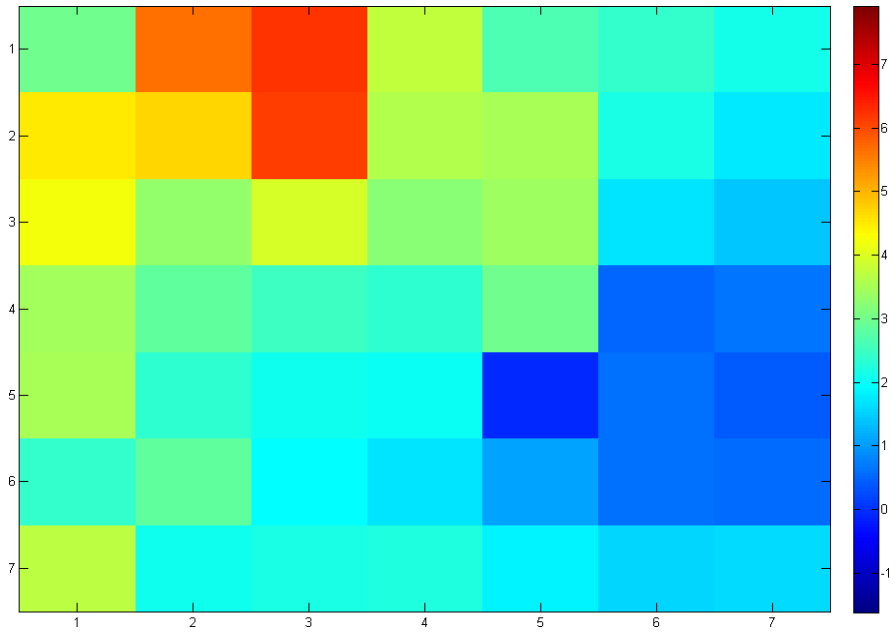

Figura 8: Promedio predicción sobre SOM

En la Tabla 2 se muestra el resultado de las predicciones alcanzadas por el NNE. Aquellos errores mayores a $0{,}5$ unidades de logP se muestran en negrita. A juzgar por los errores promedios cuadrados (*mean square error - MSE*) y los errores promedios absolutos (*mean absolute error - MAE*), podemos alegar que la red tiene un comportamiento aceptable: MSE$= 0{,}1867$ y MAE$= 0{,}3018$. No obstante proponemos tomar aquellos compuestos con los que tuvimos un error en la predicción mayor a $0{,}5$ unidades de logP, e intentar mejorar estas predicciones con la ayuda del SOM.

Para esto, analizamos primero cuáles son los nodos que pertenecen a cada uno de los compuestos con mala predicción. Luego, hacemos tender la predicción obtenida por el NNE ($M(x)$) al valor promedio de esa celda ($N(x)$) según la ecuación 4, donde x es el vector de descriptores del caso presentado. La Tabla 3 nos muestra esta información resumida con el valor de predicción corregido ($T(x)$). Podemos ver que este proceso de corregir las predicciones en sentido del promedio del SOM, disminuyó el error obtenido en la mayoría de los casos. Los errores generales logrados ahora son: MSE$= 0{,}1624$ y MAE$= 0{,}2791$.

$$T(x) = M(x) + \frac{(N(x) - M(x))}{2}. \tag{4}$$

Ind.[a]	**logP**$_{exp}$	**NNE**	**Error**	**Ind.**	**logP**$_{exp}$	**NNE**	**Error**
62	4,57	4,2767	-0,2933	288	1,21	1,1201	-0,0899
34	4,38	4,3406	-0,0394	426	2,52	1,6341	**-0,8859**
309	4,11	3,7555	-0,3545	425	1,81	1,9266	0,1166
135	2,34	3,4080	**1,0680**	171	2,34	2,4683	0,1283
67	1,77	1,7066	-0,0634	418	0,4	0,6393	0,2393
344	3,13	3,1905	0,0605	433	1,44	1,6735	0,2335
132	2,72	3,1232	0,4032	346	3,13	3,7784	**0,6484**
249	4	3,8185	-0,1815	438	0,27	0,4762	0,2062
427	2,69	2,5138	-0,1762	357	2,31	2,1065	-0,2035
341	3,44	3,4488	0,0088	350	2,06	1,9062	-0,1538
329	2,11	1,7346	-0,3754	355	4,22	4,0842	-0,1358
198	5,07	4,9692	-0,1008	22	2,37	2,4189	0,0489
279	4,14	4,5585	0,4185	105	2,14	1,9661	-0,1739
286	3,15	3,0897	-0,0603	43	2,1	2,0841	-0,0159
63	3,5	3,5943	0,0943	89	5,15	5,3265	0,1765
254	2,86	2,7136	-0,1464	120	2,77	3,4289	**0,6589**
59	0,65	0,8186	0,1686	429	3,2	1,3818	**-1,8182**
247	4,01	3,8748	-0,1352	207	2,3	3,0049	**0,7049**
129	3,21	3,0340	-0,1760	16	2,82	2,6862	-0,1338
21	1,99	2,0062	0,0162	38	2,75	2,6593	-0,0907
10	4	4,3054	0,3054	199	2,92	3,9554	**1,0354**
259	1,72	2,3476	**0,6276**	6	0,75	0,7928	0,0428
250	3,45	3,5250	0,0750	326	3,42	3,0776	-0,3424
301	-0,54	-0,3606	0,1794	7	1,24	0,7864	-0,4536
261	4,09	3,2498	**-0,8402**	178	4,61	4,6218	0,0118
197	3,58	3,4629	-0,1171	102	1,32	1,6476	0,3276
196	2,42	3,0746	**0,6546**	24	2	2,4709	0,4709
208	1,88	2,3387	0,4587	30	4,11	3,3222	**-0,7878**
192	-1,38	-1,1113	0,2687	103	2,58	2,4402	-0,1398
253	0,81	0,8381	0,0281	114	1,9	1,7202	-0,1798
399	2,19	2,3337	0,1437	227	2,99	2,6152	-0,3748
362	1,95	1,7794	-0,1706	177	3,33	3,0110	-0,3190
252	1,23	1,2533	0,0233	81	5,02	5,2585	0,2385
391	2,94	3,0198	0,0798	15	1,48	2,7586	**1,2786**
189	3,23	2,8301	-0,3999	432	1,97	2,3688	0,3988
61	3,15	2,3426	**-0,8074**	387	1,56	1,4534	-0,1066
188	1,94	1,8228	-0,1172	361	1,85	1,9972	0,1472
328	-0,66	-0,4611	0,1989	331	0,18	0,5640	0,3840
312	0,59	0,6851	0,0951	320	1,78	1,8777	0,0977
165	1,16	1,2516	0,0916	382	1,26	1,5178	0,2578
176	2,43	2,2360	-0,1940	407	-0,22	0,3407	**0,5607**
121	1,71	1,8857	0,1757	293	-0,3	-0,0447	0,2553
294	1,04	1,2158	0,1758	411	0,46	0,7713	0,3113

Tabla 2: Predicciones NNE para el conjunto de testeo.[a] Índice del compuesto según Yaffe et al. (2002)

5. CONCLUSIONES

El presente trabajo apunta a mostrar una modificación al modo clásico al cual se entrenan los métodos de predicción basados únicamente en el aprendizaje supervisado. La propuesta radica en el uso de mapas auto-organizativos como método de ajuste para compuestos mal modelados. Sin embargo, nuestro trabajo se presenta como una experimentación en avance, ya que aún restan ciertas cuestiones a resolver.

En primer término, la principal deficiencia del método presentado radica en la incapacidad de identificación automática de aquellas entidades químicas que resultaran mal modeladas por el método de predicción. Por consiguiente, actualmente estamos evaluando metodologías que permitan reconocer los compuestos cuyo valor de predicción sea poco confiable. En este punto,

Ind.	logP Exp	NNE	Error	Nodo Asoc.	$T(x)$	Error 2
135	2,34	3,4080	1,0680	46	2,8113	0,4713
259	1,72	2,3476	0,6276	42	1,4564	-0,2636
261	4,09	3,2498	-0,8402	15	3,7387	-0,3513
196	2,42	3,0746	0,6546	15	3,6511	1,2311
61	3,15	2,3426	-0,8074	29	2,9283	-0,2217
426	2,52	1,6341	-0,8859	33	0,7738	-1,7462
346	3,13	3,7784	0,6484	8	4,1494	1,0194
120	2,77	3,4289	0,6589	26	3,1998	0,4298
429	3,2	1,3818	-1,8182	5	2,0094	-1,1906
207	2,3	3,0049	0,7049	25	2,6929	0,3929
199	2,92	3,9554	1,0354	11	3,7698	0,8498
30	4,11	3,3222	-0,7878	26	3,1465	-0,9635
15	1,48	2,7586	1,2786	21	2,0825	0,6025
407	-0,22	0,3407	0,5607	41	0,4687	0,6887

Tabla 3: Predicciones mejoradas con el SOM.

encontramos que nuevamente el SOM, con los desvíos estándares de los NG, pueden darnos información valiosa al respecto. Asimismo, una vez desarrollado el método, la validación del mismo, debería probarse de manera más general. En este contexto, la ecuación 4, podría ser mejorada y tener en cuenta otros parámetros

No obstante, consideramos que la técnica propuesta resulta promisoria ya que, en similares experimentaciones con el aporte del SOM, se logra corregir los valores de predicción de baja precisión. Al mismo tiempo, el SOM provee de una herramienta de visualización muy interesante, dado que en función de la homogeneidad dentro de las celdas y la contigüidad en celdas vecinas se puede inferir si los descriptores usados son relevantes para el modelado de la propiedad.

6. AGRADECIMIENTOS

Los autores agradecen a SeCyT (UNS) por los proyectos PGI 24/N019, 24/ZN15, 24/ZN16.

REFERENCIAS

Agatonovic-Kustrin S. and Beresford R. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of Pharmaceutical and Biomedical Analysis*, 22:717–727, 2000.

Bishop C. *Pattern Recognition and Machine Learning*, volume I. Springer Science + Business Media, 2006.

Erös D., Kovesdi I., Örfi L., Takács-Novák K., Acsády G., and Kéri G. Reliability of logP predictions based on calculated molecular descriptors: a critical review. *Current Medicinal Chemistry*, 9:1819–1829, 2002.

Espinosa G., Arenas A., and Giralt F. An integrated som-fuzzy artmap neural system for the evaluation of toxicity. *Journal of Chemical Information and Computer Science*, 42:343–359, 2002.

Gama J. and Brazdil P. Cascade generalization. *Machine Learning*, 41:315–343, 2000.

Hansch C. and Leo A. *Substituent Constants for Correlation Analysis in Chemistry and Biology*,

volume 1. John Wiley, 1979.

Jónsdottir S., Jørgensen F., and Brunak S. Prediction methods and databases within chemoinformatics: emphasis on drugs and drug candidates. *Bioinformatics*, 21:2145–2160, 2005.

Johnson R. and Wichern D. *Applied Multivariate Statistical Analysis*, volume III. Prentice Hall, 1992.

Kühne R., Ebert R.U., and Schüürmann G. Model selection based on structural similarity-method descrption and application to water solubility prediction. *Journal of Chemical Information and Modeling*, 46:636–641, 2006.

Kohonen T. *Self-Organizing Maps*, volume II. Springer, 1997.

Livingstone D. The characterization of chemical structures using molecular properties. A survey. *Journal of Chemical Information and Computer Science*, 40:195–209, 2000.

Martin Y., Kofron J., and Traphagen L. Do structurally similar molecules have similar biological activity? *Journal of Medicinal Chemistry*, 45:4350–4358, 2002.

Selick H., Beresford A., and Tarbit M. The emerging importance of predictive ADME simulation in drug discovery. *Drug Discovery Today*, 7:109–116, 2002.

Sheridan R., Feuston B., Maiorov V., and Kearsley S. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *Journal of Chemical Information and Computer Science*, 44:1912–1928, 2004.

Soto A., Cecchini R., Vazquez G., and Ponzoni I. A wrapper-based feature selection method for ADMET prediction using evolutionary computing. *Lecture Notes in Computer Science*, 4973:188–199, 2008.

Soto A., Ponzoni I., and Vazquez G. Predicting physicochemical properties for drug design using clustering and neural network learning. *Brazilian Symposium on Bioinformatics*, pages 46–57, 2007.

Taskinen J. and Yliruusi J. Prediction of physicochemical properties based on neural network modelling. *Advanced Drug Delivery Reviews*, 55:1163–1183, 2003.

Tetko I. Associative neural network. *Neural Processing Letters*, 16:187–199, 2002a.

Tetko I. Neural network studies. 4. Introduction to associative neural networks. *Journal of Chemical Information and Computer Science*, 42:717–728, 2002b.

Tetko I., Bruneau P., Mewes H.W., Rohrer D., and Poda G. Can we estimate the accuracy of ADME-Tox predictions? *Drug Discovery Today*, 11:700–707, 2006.

Todeschini R. and Consonni V. *Handbook of Molecular Descriptors*, volume I. Wiley-VCH, 2000.

Willet P. Chemoinformatics - similarity and diversity in chemical libraries. *Current Opinion in Biotechnology*, 11:85–88, 2000.

Winkler D. Neural networks in ADME and toxicity prediction. *Drugs of the Future*, 29:1043–1057, 2004.

Wolpert D. Stacked generalization. *Neural Networks*, 5:241–260, 1992.

Yaffe D., Cohen Y., Espinosa G., Arenas A., and Giralt F. Fuzzy ARTMAP and back-propagation neural networks based quantitative structure - property relationships (QSPRs) for octanol-water partition coefficient of organic compounds. *Journal of Chemical Information and Computer Science*, 42:162–183, 2002.