ARTÍCULO ORIGINAL

Metodología multi-modal en relaciones cuantitativas estructura-actividad

Multi-Modal approach in quantitative structure-activity relationships studies

Lisset Cabrera-Leyva,[I] Julio Cesar Madera Quintana,[I] César R. García-Jacas,[II] Yovani Marrero-Ponce[III]

I Grupo de Investigación de Inteligencia Artificial (AIRES), Facultad de Informática, Universidad de Camagüey, Camagüey, Cuba. E-mail: lisset.cabrera@reduc.edu.cu
II Escuela de Sistemas y Computación, Pontificia Universidad Católica del Ecuador Sede Esmeraldas (PUCESE), Esmeraldas, Ecuador.
Grupo de Investigación de Bioinformática, Centro de Estudio de Matemática Computacional (CEMC), Universidad de las Ciencias Informáticas (UCI), La Habana, Cuba.
III Universidad San Francisco de Quito (USFQ), Grupo de Medicina Molecular y Traslacional (MeM&T), Colegio de Ciencias de la Salud (COCSA), Quito, Ecuador.

RESUMEN

Los estudios QSAR definidos en la literatura están basados en enfoques uni-modales, dejando de analizar conjuntos de datos que contienen distintas informaciones químicas. En esta investigación se propone aplicar por primera vez y analizar el comportamiento del enfoque multi-modal en el desarrollo de estudios QSAR. Para este fin se utilizó una base de compuestos con actividad hepatotóxica, a partir de la cual se construyeron cuatro modalidades considerando distintos descriptores moleculares basados en diversas teorías y enfoques. Se desarrollaron varios modelos usando los enfoques uni-modales y multi-modales utilizando algoritmos de clasificación reportados en la literatura e implementados en el lenguaje R. Los parámetros de cada uno de los algoritmos se optimizaron con el procedimiento "parametertuningwithrepeatedgrid-searchcross-validation", mientras la validación de dichos modelos se realizó mediante validación cruzada de 10 pliegues con 10 repeticiones. Estadísticamente se comprobó que el enfoque multimodal mejora el desempeño de los modelos predictivos comparado con algunos de los modelos derivados de los conjuntos de datos con modalidades individuales.

Palabras Clave: enfoque multi-modal, enfoque uni-modal, estudios QSAR.

ABSTRACT

The QSAR studies defined in the literature are based on uni-modal approaches and do not consider datasets with different chemical information. Thus, this research has as objective to apply and analyze the behavior of multi-modal approaches when QSAR studies are carried out. To this end, a compound dataset with hepatotoxicity activity was employed and four modalities were built considering molecular descriptors based on different mathematical theories. Also, several predictive models were developed taking into account both uni-modal and multi-modal approaches by using classification algorithms reported in the literature and implemented in R language. The parameters of these algorithms with the procedure "parameter tuning with repeated grid-search cross-validation" were optimized, while the strategy 10-fold cross-validation with 10 repetitions was used to corroborate the predictive accuracy of the models. As result of this study it can be stated that the behavior of the models based on multi-modal approach present significant differences with to those models developed from uni-modal approaches.

Key Words: multi-modal approach, uni-modal approach, QSAR studies.

INTRODUCCIÓN

Las enfermedades neoplásicas constituyenun importante problema de salud a nivel mundial. En América Latina anualmente mueren más de un millón de personas por esta enfermedad, mientras en Cuba el cáncer constituye desde el 2012 la primera causa de muerte. Específicamente el 2014 cerró con una tasa de 215.5 fallecidos por cada 100 000 habitantes.

Entre las enfermedades neoplásicas de mayor interés son aquellas asociadas con el hígado al ser este uno de los órganos imprescindibles para la vida. La hepatotoxicidad (HTX) también llamada enfermedad hepática tóxica o cáncer de hígado inducida por medicamentos se define como la lesión o daño hepático causado por la exposición a un medicamento u otros agentes no farmacológicos.[1]

Para el estudio de esta enfermedad pueden ser aplicados métodos computacionales para comprender y explicar la relación existente entre las características moleculares y su actividad o efecto en el organismo, los que son conocidos como estudios QSAR. Estos están dirigidos a encontrar buenas correlaciones entre las características o descriptores moleculares y actividades biológicas específicas para así obtener modelos con buena capacidad de predicción en nuevas entidades químicas.[1]

Hasta la fecha los métodos QSAR definidos en la literatura están basados en enfoques uni-modales, es decir, una única base compuesta por el mismo tipo de descriptor molecular (DM) [ej. descriptores topológicos] y/o calculada por un solo

software. Este enfoque tiene dos desventajas principales: 1) si el conjunto de datos solamente contiene información topológica entonces no se considera información geométrica y viceversa, y 2) si los DMs son calculados con un solo software entonces no se toman en cuenta DMs calculados con definiciones matemáticas diferentes y calculadas con otras herramientas computacionales.

Diferentes investigaciones han reportado que los enfoques multimodales mejoran el desempeño comparado con los mejores modelos derivados de los conjuntos de datos con modalidades individuales.[2] El objetivo de este estudio es aplicar por primera vez y analizar el comportamiento del enfoque multi-modal en el desarrollo de estudios QSAR con el propósito de predecir qué compuestos pueden presentar actividad hepatotóxica.

MATERIALES Y MÉTODOS

Conjunto de datos químicos

Para llevar a cabo este estudio se utilizó la base de compuestos con actividad hepatotóxica disponible en: http://padel.nus.edu.sg/software/padelddpredictor/mdels/toxicity/hepatotoxicity/20110523/. Esta base ha sido empleada en otros estudios[3] y la misma está constituida por 1087 moléculas no cogenéricas (que no pertenecen a la misma familia química)de las cuales 654 son hepatotóxicas y 433 no hepatotóxicas.

Para realizar el presente estudio se calcularon, en el conjunto de compuestos químicos considerado, distintos DMs basados en diversas teorías y enfoques, utilizando los siguientes software: DRAGON,[4] PaDEL-Descriptor[5] y QuBiLS-MIDAS.[6] Los DMs determinados por cada uno de estos programas se agruparon acorde a su definición matemática y al tipo de información química codificada en los grupos que se mencionan a continuación:

- **Otros_Dragon-Padel:** en esta base se encuentran los DMs de tipo conteo, fragmentos y huellas (fingerprints) calculados con los software DRAGON y PADEL. EstosDMs son identificadoscomo: 0D-DMs, 1D-DMs, 2D binary fingerprints y 2D frequency fingerprints del software Dragon; CDK extended fingerprint, Estate fingerprint, MACCS fingerprint, Substructure fingerprint count y Klekota-Roth fingerprint del software PaDEL.

- **2D_Dragon-Padel:** en esta base se encuentran los DMs de tipo topológico calculados con los programas DRAGON y PADEL.

- **3D_Dragon-Padel:** en esta base se encuentran los DMs de tipo geométricos calculados con los programas DRAGON y PADEL.

- **3D_QuBiLS-MIDAS:** en esta base se encuentran los DMs de tipo geométricos calculados con el software QuBiLS-MIDAS. Estos índices 3D están basados en álgebra tensorial y emplean diferentes métricas de distancia y multi-métricas para codificar información para relaciones entre dos, tres y cuatro átomos.

Como el cálculo de estos DMs conlleva a un espacio de alta dimensionalidad entonces se realizaron los siguientes pasos con el propósito de encontrar un subconjunto adecuado para cada conjunto considerado y obtener un buen desempeño por los algoritmos de clasificación:

1- Se normalizaron los rangos de cada rasgo (DMs) en el intervalo de [0-1] con el método Característica de Escala definido por Y. Marrero Ponce y colaboradores.[1]

2- Se aplicaron tres filtros para remover los rasgos con información redundante e irrelevante:
- Filtro Varianza cercana a cero: remueve aquellos rasgos (DMs) donde los valores son constantes o casi constantes. Para este fin, se utilizó la función nearZeroVar del paquete caret implementado en el lenguaje R.
- Filtro Rango inter-cuartílico (IQR): este filtro fue usado para eliminar algunos rasgos con baja variabilidad (IQR cercana a 0) porque no son capaces de discriminar a través de diferentes tipos de clases.
- Filtro Correlación: elimina aquellos rasgos con una correlación mayor que 0.9. Para ello se utilizó la función find Correlation del paquete caret implementado en el lenguaje R. (Tabla 1)

Tabla 1. Número de descriptores (rasgos) en el conjunto de compuestos químicos, antes y después del proceso de filtrado

Modalidad	Núm. Rasgos (antes)	Núm. rasgos (después)
2D_Dragon-Padel	2392	205
3D_Dragon-Padel	1003	141
otros_Dragon-Padel	7264	656
3D_ToMoCoMD	2196	390

Modelado

Con el propósito de analizar el desempeño de la capacidad predictiva de los modelos a desarrollar se usaron dos tipos de enfoques descritos a continuación (ver Figuras 1 y 2 para una representación gráfica):

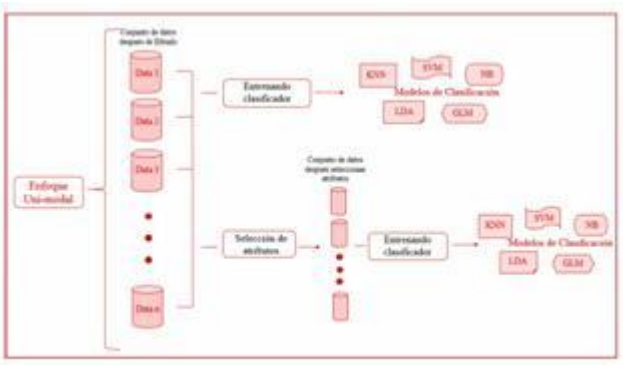

Fig. 1. Enfoque Uni-modal

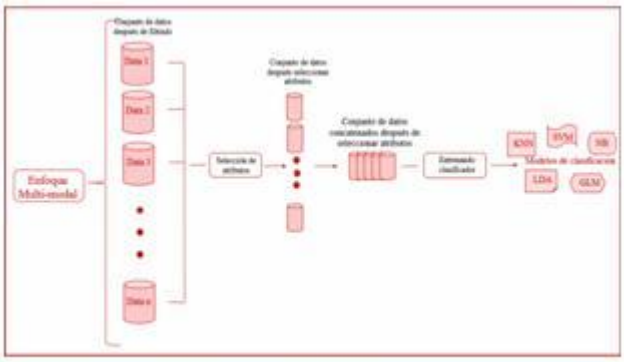

Fig. 2. Enfoque Multi-modal

- Enfoque uni-modal (enfoque tradicional): selección de rasgos y métodos de clasificación supervisada a un conjunto de datos con un mismo tipo de información (modalidad única).

- Enfoque multi-modal: selección de rasgos y métodos de clasificación supervisada a conjunto de datos con múltiples modalidades para obtener modelos predictivos.

En el enfoque uni-modal se analizó cada conjunto de datos de manera independiente en dos partes: uno sin seleccionar atributos y otro con selección de atributos. Por otra parte en el enfoque multi-modal se aplicó selección de atributos a cada conjunto de datos de forma independiente y se concatenaron los DMs (rasgos) resultantes de esa selección. La técnica de selección utilizada en ambos casos es el algoritmo CFS correspondiente al paquete caret implementado en el lenguaje R.

Para el desarrollo de los modelos QSAR se utilizaron los siguientes algoritmos de clasificación:

- Algoritmo K-vecinos más cercanos: se optimizó el parámetro . Se utilizó la función knn del paquete caret implementado en el lenguaje R. Los mejores modelos fueron obtenidos con .

- Algoritmo Máquina de soporte vectorial: se usaron las funciones kernel lineal, radial y polinomial. También se optimizaron los parámetros sigma(?) con los valores 0.01, 0.1, 0.2, 1 y el parámetro C con los valores 0.001,0.01, 0.1, 1, 10, 100,1000. Se utilizaron las funciones svmLinear, svmRadial, svmPoly del paquete carety del paquete kernlab implementado en el lenguaje R. Los mejores modelos fueron obtenidos con sigma = 0.01 y C = 0.01 y 0.001.

- Algoritmo Redes bayesianas: se utilizó la función nbdel paquete caret y del paquete kernlab implementado en el lenguaje R.

- Algoritmo Análisis discriminante lineal: se utilizó la función lda del paquete caret implementado en el lenguaje R.

- Algoritmo Modelos lineal generalizados: se utilizó la función glm del paquete caret implementado en el lenguaje R.

Para validar los modelos QSAR desarrollados se utilizó el procedimiento 10-fold cross-validation con 10 repeticiones.[7] Para optimizar los parámetros de los algoritmos de clasificación fue empleado el algoritmo

parametertuningwithrepeatedgrid-searchcross-validation,[7] el mismo devuelve el parámetro optimizado con mínimo error medio de validación cruzada.

Para evaluar la calidad de los algoritmos de clasificación se utilizaron las medidas siguientes:

- El área debajo de la curva ROC (AUC): (2)

- Sensibilidad: mide la proporción de los positivos que son correctamente identificados: (3)

- Especificidad: mide la proporción de los negativos que son identificados correctamente: (4)

RESULTADOS Y DISCUSIÓN

Para evaluar los resultados alcanzados se compararon las medidas SEN, SPE y AUC. Para el procesamiento estadístico de todos los resultados experimentales se usó el KEEL[8] en su versión 3.0. Los resultados obtenidos en cada una de las bases con cada uno de los algoritmos de clasificación son mostrados en las tabla 2 y en las figuras 3 y 4.

Tabla 2. AUC de los diferentes métodos de clasificación por cada conjunto de dato

	knn	svmL	svmR	svmP	nb	lda	glm
2D Dragon-Padel	0.71	0.70	0.74	0.72	0.72	0.70	0.70
3D Dragon-Padel	0.71	0.68	0.73	0.73	0.71	0.68	0.68
Otros Dragon-Padel	0.71	0.73	0.74	0.73	0.71	0.72	0.72
3D ToMoCoMD	0.74	0.76	0.78	0.77	0.79	0.74	0.74
Data_concat	0.73	**0.78**	**0.78**	**0.79**	0.77	0.74	**0.74**

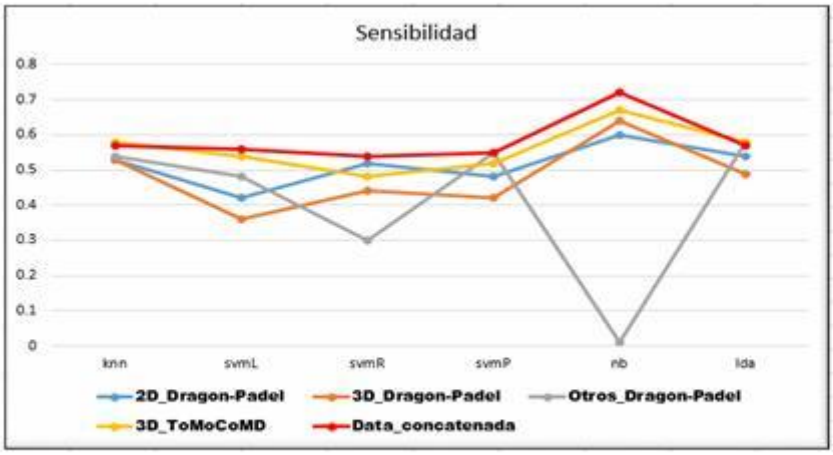

Fig. 3. SEN de los diferentes métodos de clasificación por cada conjunto de dato

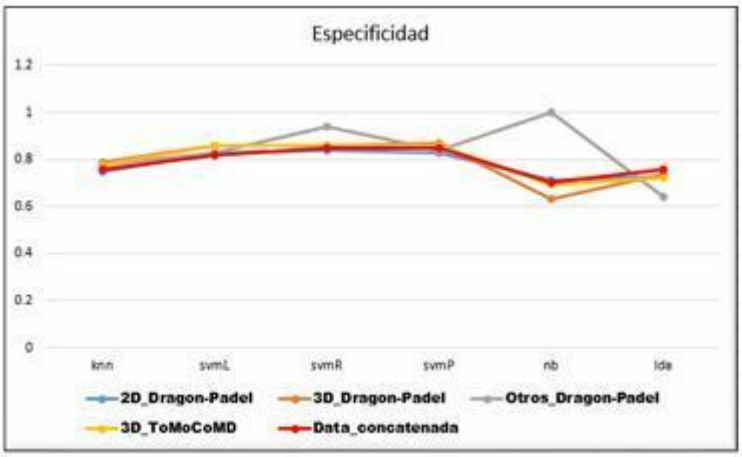

Fig. 4. SPE de los diferentes métodos declasificación por cada conjunto de dato

Particularmente se puede ver que de los 5 conjuntos de datos analizados los conjuntos 3D_ToMoCoMD (una modalidad del enfoque Uni-modal) y Data_concat (enfoque Multi-modal) tienen un comportamiento similar entre sí y superior a los restantes modalidades.

Para el análisis estadístico de los resultados se utilizaron las técnicas de prueba de hipótesis.[9] Para comparaciones múltiples se utiliza el test de Friedman para detectar diferencias estadísticas globales entre un grupo de resultados. Se emplea además la prueba Wilcoxon para determinar dife- rencias estadísticas particulares entre los enfoques considerados.

En la tabla 3 se puede observar que el mejor ranking para la medida AUC es obtenido por el enfoque multi-modal (dataSet_Concatenated).El p-value calculado por el test de Friedman es 0.000176. De esta forma se confirma el objetivo de la presente investigación al ser el enfoque multi-modal el de mejor comportamiento.

Tabla 3. Rango promedio obtenido por cada método (pruebas de Friedman)

Algorithm	Ranking
2D-Dragon-Padel	4
3D-Dragon-Padel	4.4286
otros-Dragon-Padel	3.5714
3D-ToMoCoMD	1.7143
dataSet-Concatenated	1.2857

Por otro lado en las figuras 5 y 6 se muestran los resultados de la prueba de Wilcoxon para el enfoque multi-modal con respecto a cada modalidad individual con respecto a la medida AUC.

VS	R^+	R^-	Exact P-value	Asymptotic P-value
2D-Dragon-Padel	28.0	0.0	0.015626	0.014248
3D-Dragon-Padel	28.0	0.0	0.015626	0.014248
otros-Dragon-Padel	28.0	0.0	0.015626	0.014248
3D-ToMoCoMD	19.0	9.0	≥ 0.2	0.352542

Fig. 5. Resultados obtenidos por las pruebas de Wilcoxon para el enfoque Multi-modal

	(1)	(2)	(3)	(4)	(5)
2D-Dragon-Padel (1)	-	•		○	○
3D-Dragon-Padel (2)		-	○	○	○
otros-Dragon-Padel (3)			-	○	○
3D-ToMoCoMD (4)	•	•	•	-	
dataSet-Concatenated (5)	•	•	•		-

Fig. 6. Resumen del test de Wilcoxon

- • los métodos de las filas mejoran al método de las columnas.
- ○ los métodos de las columnas mejoran al método de las filas.

Como se puede ver las modalidades individuales 2D_Dragon-Padel, 3D_ Dragon-Padel y otros_ Dragon-Padel son significativamente inferiores al enfoque multi-modal, mientras la modalidad individual 3D-ToMoCoMD no arrojó diferencias significativas con respecto al enfoque multi-modal, aunque este último si tiene un mejor comportamiento.

CONCLUSIONES

En este estudio se analizó el comportamiento del enfoque Multi-modal en el desarrollo de los estudios QSAR (análisis desarrollado por primera vez) para identificar los compuestos de acuerdo con las actividades biológicas, utilizando la base hepatotóxica.

Se demostró estadísticamente que el enfoque multimodal en los estudios QSAR mejora el desempeño comparado con algunos los modelos derivados de los conjuntos de datos con modalidades individuales, con otras modalidades individuales como por ejemplo 3D-ToMoCoMD mostró un comportamiento similar.

REFERENCIAS BIBLIOGRÁFICAS

1. Marrero Y, Santiago O-M, López Y-M, Barigye S-J, Torrens F. "Derivatives in discrete mathematics: a novel graph-theoretical invariant for generating new 2/3D molecular descriptors. I. Theory and QSPR application". Journal of computer-aided molecular design. vol. 26, pp. 1229-1246, 2012.

2. Ray B, Henaff M, Ma S, Efstathiadis E, Peskin E-R, Picone M, Poli T, Aliferis C-F, Statnikov A. "Information content and analysis methods for Multi-Modal High-Throughput Biomedical Data". Scientific reports. vol. 4, 2014.

3. Liew C-Y, Lim Y-C, Yap C-W. "Mixed learning algorithms and features ensemble in hepatotoxicity prediction". Journal of computer-aided molecular design. vol. 25, pp. 855-871, 2011.

4. Mauri A, Consonni V, Pavan M, Todeschini R. "Dragon software: An easy approach to molecular descriptor calculations". Match. vol. 56, pp. 237-248, 2006.

5. Yap CW. "PaDEL?descriptor: An open source software to calculate molecular descriptors and fingerprints". Journal of computational chemistry. vol. 32, pp. 1466-1474, 2011.

6. García C. R, Marrero Y, Ponce L, Barigye S. J, Valdés J. R, Contreras E. "QuBiLS?MIDAS: A parallel free?software for molecular descriptors computation based on multilinear algebraic maps". Journal of computational chemistry. vol. 35, pp. 1395-1409, 2014.

7. Krstajic D, Buturovic L. J, Leahy D. E, Thomas S. "Cross-validation pitfalls when selecting and assessing regression and classification models". Journal of cheminformatics. vol. 6, pp. 1-15, 2014.

8. Alcalá 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. vol. 17, pp. 255-287, 2010.

9. Sheskin D. "Handbook of parametric and nonparametric statistical procedures, chapman & hall". presented at the CRC, 2003.

Recibido: 22 de marzo de 2016.
Aprobado: 12 de mayo de 2016.