

Multiclasificadores basados en aprendizaje automático como herramienta para la evaluación del perfil neurotóxico de líquidos iónicos

R. Domínguez-López¹, L.C. Avila-de-la-Cruz¹, F. Borges², M.N.D.S. Cordeiro³, C. Paz-y-Miño⁴, E. Tejera⁴, Y. Pérez-Castillo⁵, A. Sánchez-Rodríguez⁶, Y. Perera-Sardiña⁷, M. Cruz-Monteagudo^{2,4*}

¹Dpto. Computación, Univ. de Las Tunas, Cuba. ²CIQUP ³REQUIMTE Dpto. Quím. e Bioquímica, Fac. Ciências, Univ. do Porto, Portugal. ⁴IIB, Univ. de Las Américas, Quito, Ecuador. ⁵Sec. Físico Quím. y Matemáticas, Dpto. Quím., Univ. Técnica Particular de Loja, Ecuador. ⁶Dpto. Ciencias Nat., Univ. Técnica Particular de Loja, Ecuador. ⁷Dpto. Ciencias Quím., Fac. Ciencias Exactas, Univ. Andres Bello, Santiago de Chile, Chile.

Machine learning based multiclassifiers as a neurotoxicity estimation tool for ionic liquids

Multiclassificadors basats en l'aprenentatge automàtic com a eina per a l'avaluació del perfil neurotòxic de líquids iònics

RECEIVED: 11 MAY 2016. REVISED: 6 JUNE 2016. ACCEPTED: 6 JUNE 2016

SUMMARY

Ionic liquids (ILs) possess a unique physicochemical profile providing a wide range of applications. Their almost limitless structural possibilities allow the design of task-specific ILs. However, their “greenness,” specifically their claimed relative nontoxicity has been frequently questioned, hindering their REACH registration processes and, so, their final application. Since the most of ionic liquids has yet to be synthesized, the development of chemoinformatic tools allowing the efficient profiling of the hazardous potential of these compounds becomes critical. In this sense, the combined use of multiple base classifiers (ensembles or multiclassifiers) have proved to overcome the classification performance limitations associated to the use of single classifiers. In the present work we report two ensembles models with good predictive capabilities in a validation set of ionic liquids never used in the learning process. These models were obtained as part of Quantitative Structure Activity Relationship's studies (QSAR) applied to the characterization of neurotoxic profile of ionic liquids based on its inhibition of the Acetyl cholinesterase enzyme (AChE) as neurotoxicity indicator. The results obtained show that one can expect that at least 96% of a set of news

ionic liquids can be correctly classified using these ensembles models. Consequently, these chemoinformatics models provides efficient decision making tools in the design and development of new “green” ionic liquids.

Key words: Acetylcholinesterase; ionic liquids; QSAR; multiclassifiers; diversity measures; AdaboostM.1; stacking.

RESUMEN

Los líquidos iónicos poseen un perfil físicoquímico único, el cual los provee de un amplio rango de aplicaciones. Su variabilidad estructural casi ilimitada permite su diseño para tareas específicas. Sin embargo, su sustentabilidad, específicamente su seguridad desde el punto de vista toxicológico, ha sido frecuentemente cuestionada. Este último aspecto limita significativamente el cumplimiento de las regulaciones establecidas por la Unión Europea para el registro,

*Corresponding author: gmailkelcm@yahoo.es

evaluación, autorización y restricción de compuestos químicos (REACH), así como su aplicación final. Debido a que la mayoría de los líquidos iónicos no han sido sintetizados, se hace evidente la importancia del desarrollo de herramientas quimioinformáticas que, de forma eficiente, permitan evaluar el potencial toxicológico de estos compuestos. En este sentido, el uso combinado de múltiples clasificadores ha demostrado superar las limitaciones de desempeño asociadas al uso de clasificadores individuales. En el presente trabajo fueron evaluadas varias estrategias alternativas de multclasificadores basados en técnicas de aprendizaje automático supervisado, como herramientas para la evaluación del perfil neurotóxico de líquidos iónicos basado en la inhibición de la enzima acetilcolinesterasa, como indicador de neurotoxicidad. Se obtuvieron dos multclasificadores con una alta capacidad predictiva sobre un conjunto de validación externa (no utilizado en el proceso de aprendizaje de los modelos). De acuerdo a los resultados obtenidos el 96% de un conjunto de nuevos líquidos iónicos podrá ser correctamente clasificado con la utilización de estos multclasificadores, los cuales constituyen herramientas de toma de decisión útiles en el campo del diseño y desarrollo de nuevos líquidos iónicos sustentables.

Palabras clave: Acetilcolinesterasa; líquidos iónicos; QSAR; multclasificadores; medidas de diversidad; AdaboostM.1; Stacking.

RESUM

Els líquids iònics posseeixen un perfil fisicoquímic únic, el qual els proveeix d'un ampli rang d'aplicacions. La seva variabilitat estructural quasi il·limitada permet el seu disseny per a tasques específiques. No obstant, la seva sostenibilitat, específicament la seva seguretat des del punt de vista toxicològic, ha estat freqüentment qüestionat. Aquest darrer limita significativament el compliment de les regulacions establertes per la Unió Europea pel registre, avaluació, autorització i restricció de compostos químics (REACH), així com la seva aplicació final. Degut a que la majoria dels líquids iònics no han estat sintetitzats, es fa evident la importància del desenvolupament d'eines quimioinformàtiques que, de manera eficient, permeten avaluar el potencial toxicològic d'aquests compostos. En aquest sentit, l'ús combinat de múltiples classificadors ha demostrat superar les limitacions d'acompliment associades l'ús de classificadors individuals. En el present treball van ser avaluades diverses estratègies alternatives de multclasificadors basats en tècniques d'aprenentatge automàtic supervisat, com a eines per a l'avaluació del perfil neurotóxic de líquids iònics basat en la inhibició de l'enzim acetilcolinesterasa, com a indicador de neurotoxicitat. Es van obtenir 2 multclasificadors amb una alta capacitat predictiva sobre un conjunt de validació externa (no utilitzat en el procés d'aprenentatge dels models). D'acord als resultats obtinguts el 96% d'un conjunt de nous líquids iònics

podrà ser correctament classificat amb la utilització d'aquests multclasificadors, els quals constitueixen eines de presa de decisió útils en el camp del disseny i desenvolupament de nous líquids iònics sostenibles.

Paraules clau: Acetilcolinesterasa; líquids iònics; QSAR; multclasificadors; mesures de diversitat; AdaboostM.1; Stacking.

INTRODUCCIÓN

Los líquidos iónicos (LI) constituyen un área de las ciencias químicas que actualmente atrae la atención entre la comunidad científica. Las investigaciones en torno a ese campo están motivadas por las características físico-químicas únicas de estos compuestos que proveen un amplio rango de aplicaciones¹. El número de posibles combinaciones de aniones y cationes que pudieran componer un LI está por el orden de los 10^{18} ². Los LI son denominados “verdes” o sustentables y la razón por la que son llamados de esta forma consiste generalmente en tres argumentos: i) poseen una presión de vapor muy baja que proporciona una reducción de la exposición por inhalación en el personal que trabaja con él en comparación con solventes moleculares convencionales; ii) baja inflamabilidad, que reduce significativamente el riesgo de oxidaciones exotérmicas rápidas y iii) son relativamente no tóxicos³⁻⁵.

Existen investigadores que están de acuerdo con las primeras características pero otros expresan su desacuerdo con la última característica y la justificación de su discrepancia se encuentra plasmada en varios artículos⁶. Un creciente número de estudios, que analizan el riesgo de diferentes LI, han sido realizados y determinados en diferentes sistemas de pruebas biológicas⁷⁻⁹.

A partir de estas investigaciones se pudo demostrar que los LI son potencialmente peligrosos tanto para el hombre como para el medioambiente. Como resultado fueron identificados LI con niveles que van desde bajo peligro potencial hasta aquellos que son potencialmente muy tóxicos. Dentro del sistema de ensayos utilizadas en estas investigaciones se encuentra la evaluación de la inhibición de la enzima Acetilcolinesterasa (AChE)” realizado a partir de la enzima purificada proveniente de la manta eléctrica “*Electrophorus electricus*”, apoyados en estudios QSAR. Dicha enzima es parte esencial del sistema nervioso central de los organismos complejos y por tanto puede ser usada como un indicador del potencial neurotóxico de aquellas sustancias capaces de inhibirla (10). Los resultados de este estudio permitieron determinar que los LI pueden inhibir la enzima y por consecuencia provocar efectos no deseados en los organismos afectados. Los resultados de los ensayos “*in vivo*” son un indicador de que la AChE de otros organismos también puede ser inhibida¹⁰. Como resultado de las investigaciones sobre el efecto inhibitorio de los LI sobre la AChE ha sido generado un volumen de registros donde se describen las características

estructurales de los LI ya sintetizados y su actividad sobre la AChE. A partir de que la gran mayoría de LI aún no han sido sintetizados es necesario desarrollar métodos capaces de predecir el peligro potencial de LI desconocidos con el objetivo de facilitar el diseño de nuevos materiales seguros y reducir la necesidad de realizar la síntesis de LI basados en el método de prueba y error, lo que permite disminuir los costos y el tiempo de su producción.

El uso de herramientas computacionales ha demostrado ser un factor fundamental en investigaciones realizadas en múltiples especialidades debido a que ha sido utilizada como herramienta para realizar simulaciones matemáticas y de otras índoles con el objetivo de realizar experimentos más certeros y cercanos a la realidad. La computación, como herramienta, complementa la realización de estudios que utilizan simulaciones o procesamiento de grandes volúmenes de datos que de forma manual sería inviable realizar. El aprendizaje automático es un área de la inteligencia artificial que ha demostrado ser de gran utilidad para el procesamiento de grandes volúmenes de información en las ciencias biológicas y químicas. En este sentido, se han reportado varios artículos relacionados con de herramientas computacionales para la predicción de la toxicidad de los LI¹¹⁻²². Específicamente, la predicción del potencial inhibitorio de los LI sobre la enzima AChE, como un problema de clasificación ha sido poco explorada. En este sentido los resultados reportados se limitan a la utilización de técnicas de clasificación o regresión, con un escaso reporte del uso de multclasificadores. Basados en lo anteriormente expuesto proponemos en este trabajo la identificación de varios modelos basados en multclasificadores que permitan predecir eficientemente la capacidad inhibitoria de LI sobre la enzima AChE y de esta forma poder estimar su potencial neurotóxico.

MATERIALES Y MÉTODOS

Para el desarrollo del trabajo se aplicó un enfoque de minería de datos. Se aplicaron las etapas del proceso de extracción del conocimiento en base de datos, con el objetivo de encontrar modelos que permitan la identificación de patrones a partir de la información actual y con su utilización poder predecir el peligro potencial de LI desconocidos.

Recolección y estandarización de los datos

Los datos utilizados fueron obtenidos de la Base de datos de efectos biológicos de LI pertenecientes a la UFT/Merck^{10,11,23}. Tropsha y colaboradores²⁴ proponen una metodología para preprocesar la información relacionada con la estructura de moléculas químicas que serán objeto de la realización de estudios QSAR. Esta metodología propone eliminar todas aquellas moléculas que poseen elementos metálicos u organometálicos, de manera que sean descartadas aquellas moléculas que contienen elementos que no son manejados por las herramientas para el cálculo de descriptores moleculares, así como elementos poco representados que pu-

dieran constituir ruidos en la fase de aprendizaje. Los elementos poco representados son determinados con la utilización del software JChem for Excel²⁵.

La estructura química de cada molécula, que se encuentra codificada mediante el correspondiente código SMILES (*Simplified Molecular Input Line Entry Specification*), debe ser estandarizada debido a que, en ocasiones, un mismo grupo funcional puede ser representado por diferentes patrones estructurales en un conjunto de datos dado. Para los estudios QSAR esta situación conduce a serios problemas debido a que los descriptores moleculares calculados a partir de estas representaciones diferentes del mismo grupo funcional pueden ser significativamente distintos²⁴.

Para la estandarización estructural de las moléculas se utilizó la herramienta ChemAxon's Standarizer²⁶. Seguidamente se procedió a identificar duplicados de una molécula con el objetivo de removerlas. Los duplicados pueden afectar la frecuencia observada de una molécula y la distribución de los compuestos acorde a su similitud estructural. Como consecuencia los duplicados deben ser eliminados antes de realizar cualquier estudio de modelación. Para la eliminación de duplicados se utilizó el software EdiSDF²⁷. A continuación se procedió a realizar el cálculo de los descriptores moleculares mediante la utilización de la herramienta ISIDA^{27,28}, lo que permitió la obtención de un vector numérico que representa la ocurrencia de fragmentos presentes en la molécula tal y como se describe en²⁹. Finalmente se evaluaron varios clasificadores y multclasificadores implementados en la herramienta de minería de datos Weka³⁰.

Diseño del experimento

El conjunto de datos resultante recoge la información de 261 moléculas, donde cada molécula está representada por 693 descriptores moleculares. Las moléculas se clasifican en dos grupos: 226 LI con una capacidad inhibitoria sobre la enzima AChE de mediana a muy alta y clasificados como potencialmente neurotóxicos y 35 LI con una baja capacidad inhibitoria sobre la enzima AChE y clasificados como potencialmente no neurotóxicos o seguros. Para el diseño del experimento de clasificación mediante la aplicación de la metodología QSAR, se seleccionó el 75% de las instancias de cada clase para ser utilizada como conjunto de entrenamiento y se reservó el 25% restante de cada clase para evaluar la capacidad de generalización de los modelos obtenidos (conjunto de evaluación externa). Consecuentemente, el conjunto de entrenamiento inicial queda conformado por 169 LI clasificados como potencialmente neurotóxicos y 25 LI clasificados como potencialmente no neurotóxicos o seguros. Por su parte el conjunto de evaluación externa queda conformado por 57 LI clasificados como potencialmente neurotóxicos y 10 LI clasificados como potencialmente no neurotóxicos o seguros. En este caso se empleó la opción *Create Subset / Random Sampling* implementada en el software STATISTICA 8.0³¹ para realizar un muestreo aleatorio y de esta forma conformar el conjunto de datos de entrenamiento y el conjunto de datos de prueba.

El conjunto de datos a utilizar para el entrenamiento posee dos características que hacen necesario un pre-procesamiento adicional. El conjunto de datos poseen una alta dimensionalidad con respecto a la cantidad de instancias que posee, lo que propicia problemas de sobreajuste en la clasificación y además presenta un desbalance entre la cantidad de instancias que conforman las clases en las que se clasifican estos compuestos, presentado un índice de desbalance (IR) = 6.45, el cual puede ser determinado mediante la razón entre la ocurrencia de las instancias pertenecientes a la clase mayoritaria (226) y la ocurrencia de las instancias pertenecientes a la clase minoritaria³⁵. La realización de un entrenamiento con tal conjunto de datos provoca que el criterio de los clasificadores sea parcializado a favor de las instancias más representadas³².

Selección de atributos, balance de clases y modelación

Los softwares para el cálculo de descriptores moleculares generan de cientos a miles de estos descriptores. En muchas ocasiones una gran parte de estos descriptores es información redundante y poco útil en el proceso de extracción del conocimiento. Por tal razón se hace necesario seleccionar aquellos atributos que son relevantes para el problema en cuestión y de esta forma se evita el riesgo de que los clasificadores se vean afectados por el sobreajuste. La selección de atributos relevantes, en la actualidad, se realiza mediante 3 enfoques: i) algoritmos filtros, ii) algoritmos envolventes, iii) algoritmos híbridos. En³³ se explican las características de los algoritmos que pertenecen a estas tres categorías.

Para la selección de aquellos rasgos relevantes para la clase se utilizó el software mRMR³⁴, el cual implementa una selección de atributos basada en el criterio de máxima dependencia estadística basado en información mutua. El software es capaz de encontrar un subconjunto de atributos que poseen máxima relevancia con respecto al atributo dependiente (clase) y a la vez mínima redundancia entre ellos.

La configuración de los parámetros para alcanzar estos resultados se logró especificando el tamaño del subconjunto a obtener, el criterio de selección, el valor de corte de discretización (0.0, 0.5, 1.0) y el método de selección para la combinación de relevancia y redundancia. Los parámetros específicos fueron: 50 atributos como tamaño del subconjunto, mRMR como criterio de selección, 0.0 como valor de corte de discretización (sin discretización) y MIQ como método de combinación de relevancia y redundancia.

Una vez realizada la selección de los atributos relevantes se procede a realizar el balance de las clases con el objetivo de modificar la distribución de estas y garantizar que los resultados de los clasificadores no sean influenciados por el desbalance³⁵. Como resultado, estos problemas han sido identificados como desafíos actuales de la minería de datos. Los algoritmos de clasificación con frecuencia muestran un alto desempeño de clasificación respecto a la clase mayoritaria mientras lo opuesto ocurre en la clase menos representada.

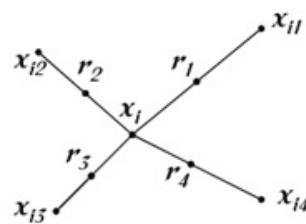


Figura 1. Proceso de generación de instancias sintéticas.

En conjuntos de entrenamientos desbalanceados la información de interés se encuentra en la clase minoritaria, muchos clasificadores consideran los datos poco representados como ruido o rarezas³⁵. En este trabajo se utilizó una técnica bien conocida y ampliamente usada para balancear las clases llamada *Synthetic Minority Oversampling Technique* (SMOTE)³⁶. SMOTE utiliza cada instancia de la clase minoritaria e introduce instancias sintéticas en el segmento que une a una/todas las k instancias más cercanas pertenecientes a la clase minoritaria. En la Figura 1, X_i representan la instancia seleccionada, X_{i1}, \dots, X_{i4} son los vecinos más cercanos seleccionados y r_1, \dots, r_4 las instancias sintéticas creadas por una interpolación aleatoria. En dependencia de la cantidad de nuevas instancias requeridas, los vecinos utilizados que pertenecen a los k vecinos más cercanos son seleccionados aleatoriamente. Para la aplicación de este algoritmo se utilizó la implementación que está presente en la herramienta Weka³⁷. Una vez obtenido el conjunto de datos de entrenamiento definitivo se procede a obtener modelos basados en multclasificadores. Para la obtención de los modelos, en una primera etapa, se emplearon ocho clasificadores, los cuales se describen en la Tabla 1.

Tabla 1. Clasificadores evaluados en la primera etapa.

Grupo	Clasificador	ID
Árboles de decisión	<i>ExtraTree</i>	ET
Red bayesiana	<i>NaiveBayes</i>	NB
Basado en instancias	<i>KStar</i>	K
Reglas	<i>Jrip</i>	J
Análisis discriminante	<i>FDLA</i>	AD
Regresión logística	<i>Logistic</i>	L
Red neuronal	<i>Multilayer Perceptron</i>	MLP
Máquina de soporte vectorial	<i>SMO</i>	SMO

A continuación se determinó el clasificador individual que mejor desempeño presentó para ser utilizado como mecanismo de comparación en la evaluación de los multclasificadores a utilizar en la experimentación. Este será utilizado para descartar aquellos modelos multclasificadores que presenten un desempeño igual o inferior. En una segunda etapa se hace necesario obtener modelos multclasificadores con la utilización de varos esquemas de agregación (*Bagging*, *Boosting*, *Stacking* y *Vote*), todos implementados en la herramienta Weka.

*Bagging*³⁸ y *Boosting*³⁹ son algoritmos que garantizan diversidad en su aprendizaje debido a la estrategia utilizada para realizar el entrenamiento. Bagging divide el conjunto de entrenamiento en varios subconjuntos, donde la selección de las instancias que conforman cada subconjunto de casos se realiza de forma aleatoria con reemplazo, de forma tal que una instancia puede formar parte en más de un subconjunto de entrenamiento. Este método utiliza un clasificador base para entrenarlo con cada subconjunto de datos generados y el método de combinación de las salidas emitidas es mediante el voto mayoritario. Por su parte, *Boosting* utiliza el mismo principio de creación de subconjuntos de entrenamiento que *Bagging* y la utilización de un solo clasificador, pero la diferencia radica en que a medida que se realiza el entrenamiento se garantiza que las instancias mal clasificadas en un subconjunto pertenezcan al próximo conjunto de entrenamiento a utilizar, de forma que permite al clasificador concentrarse en las instancias que son difíciles de clasificar. El método de combinación de las salidas generadas es mediante el voto mayoritario. En este trabajo se utilizó el algoritmo *Adaboost.M.1* como multclasificador *Boosting*. *Stacking*⁴⁰ y *Vote*⁴¹. A diferencia de los algoritmos mencionados con anterioridad, este no garantiza la diversidad en su aprendizaje, por lo que es responsabilidad del investigador garantizarla apoyándose en los criterios reportados en la literatura dedicada al tema.

Estos algoritmos, para la emisión de un criterio sobre la clase de una instancia, se basan en las salidas emitidas por los clasificadores utilizados en la construcción del modelo. En⁴² se especifican los mecanismos para garantizar "diversidad" en la combinación de clasificadores a utilizar en la construcción de un multclasificador, denominados "medidas de diversidad". En la realización de este trabajo se utilizaron cinco medidas de diversidad: i) coeficiente de correlación (p), ii) estadístico Q, iii) medida de diferencia (D), iiiii) medida de doble fallo (DF) y iiiiii) razón entre D y DF denominado como (R) las cuales son descritas en^{42,43}. Para la determinación de estas medidas de diversidad en⁴² se explica el procedimiento a realizar para calcularlas. Primeramente se determinaron las posibles combinaciones de clasificadores a utilizar, mediante la aplicación de la combinatoria mostrada en (2). La determinación de las combinaciones de clasificadores más "diversos" se realizó mediante el cálculo de las medidas de diversidad sobre las 248 posibles combinaciones de clasificadores a utilizar.

$$\text{comb} = \binom{8}{2} + \binom{8}{3} + \dots + \binom{8}{8} = 248 \quad (\text{Eq.1})$$

Esta operación permitió determinar que un total de 32 combinaciones mostraban mayor diversidad ante las restantes 216 combinaciones. Las combinaciones diversas se organizaron de forma tal que permitiera realizar una experimentación con aquellas combinaciones diversas en las cinco medidas calculadas y otra experimentación con aquellas combinaciones diversas solo en cuatro de las cinco medidas utilizadas.

Una vez en este punto se procede a continuar la experimentación. El algoritmo *Stacking* utiliza como método de combinación de las salidas de los clasificadores un clasificador, llamado "metaclasificador", que se entrena con un conjunto de instancias que el algoritmo genera, donde los componentes de cada instancia es la salida que emite cada clasificador individual utilizado en la construcción del modelo y donde el atributo dependiente continua siendo la clase original. En este caso se utilizó como metaclasificador el clasificador individual que mejor desempeño presentó en la evaluación individual realizada. El algoritmo *Vote* propone cinco métodos de combinación de las salidas de los clasificadores individuales, tales como: i) voto mayoritario, ii) promedio de probabilidades, iii) máximo de probabilidades, iiiii) mínimo de probabilidades y iiiiii) mediana de probabilidades. En este caso se utilizaron los métodos: voto mayoritario y promedio de probabilidades, métodos utilizados en la solución de un trabajo similar⁴³. La experimentación fue realizada sobre cada uno de los dos grupos de combinaciones de clasificadores determinados anteriormente como diversos. Finalmente fueron evaluados los modelos obtenidos e identificados aquellos modelos que superaron el desempeño del mejor clasificador individual. Esto demuestra que se obtuvo una mejora en el desempeño respecto a los desempeños obtenidos con los clasificadores individuales evaluados. Para la evaluación del desempeño de un clasificador se deben tener en cuenta la existencia de métricas que permiten evaluar cuan bien se comporta un modelo frente a una situación desconocida, es decir, permite evaluar la capacidad de generalización de esos modelos. Existen diversas métricas tales como la exactitud o *Accuracy* (ACC), el estadístico *Kappa* (KAPPA), la razón de verdaderos positivos (TP_{rate}), la razón de verdaderos negativos (TN_{rate}), la precisión o *Precision*, la recuperación o *Recall*, F y el área bajo la curva (AUC)^{32, 35, 44-47}. En este trabajo se emplearon métricas que miden el desempeño global de los clasificadores tales como ACC, AUC y KAPPA.

RESULTADOS Y DISCUSIÓN

Luego de aplicar las indicaciones propuestas por Tropsha y colaboradores²⁴ fueron identificadas las moléculas que poseían los siguientes elementos poco representados en el conjunto de entrenamiento: Sodio (Na), Litio (Li), Potasio (K), Hierro (Fe) y Antimonio (Sb). Las moléculas que presentaron estos compuestos fueron eliminadas del conjunto de datos de forma manual, eliminándose así 28 compuestos que podrían introducir ruido a la hora del entrenamiento de los modelos. A continuación fue utilizado el módulo Standarizer del software ChemAxon para realizar la estandarización de las moléculas. Mediante el empleo del programa EdiSDF se pudo comprobar que no existen moléculas duplicadas en el conjunto de datos. A continuación se utilizó el software ISIDA como herramienta para el cálculo de los descriptores moleculares a partir de los cuales se procedió a realizar la

experimentación. Como resultado se obtuvieron un total de 693 descriptores moleculares por cada instancia del conjunto de datos.

Selección de atributos, balance de datos y modelación

Dado que el conjunto de datos obtenido para realizar el entrenamiento posee 693 atributos y solo 195 instancias existe la posibilidad de que los clasificadores sufran de sobreajuste. Como consecuencia el clasificador será tan específico que en lugar de “aprender” de los casos “memorizará” los mismos y por tanto no será capaz de generalizar el patrón identificado para nuevos casos. Para llevar a cabo la selección de atributos se utilizó el software mRMR, que permitió la obtención de un subconjunto de atributos que son los más relevantes respecto al atributo dependiente y que poseen la menor redundancia posible entre ellos.

Luego de aplicada la configuración especificada y de ejecutado el algoritmo de selección de atributos se identificaron 50 atributos que permitieron maximizar la diferencia entre la relevancia de los atributos y la redundancia entre ellos.

A continuación se procedió a realizar el balance entre las clases. Para esto se utilizó el algoritmo SMO-TE, incluido en la herramienta Weka, el cual permitió la generación de 143 instancias artificiales pertenecientes a la clase minoritaria del conjunto de datos de entrenamiento. Consecuentemente, el conjunto de entrenamiento final queda conformado por 169 LI clasificados como potencialmente neurotóxicos y 168 LI clasificados como potencialmente no neurotóxicos o seguros (25 LI potencialmente neurotóxicos + 143 instancias artificiales generadas con SMOTE).

Una vez que se logra balancear los datos en el conjunto de entrenamiento se está en condiciones de proceder con la obtención de los modelos de clasificación. Como se explicó con anterioridad se realizó una evaluación con ocho clasificadores individuales (Ver Tabla 1). En la Tabla 2 se muestra el desempeño alcanzado por cada clasificador individual en los conjuntos de entrenamiento y de evaluación externa. Como se puede apreciar el clasificador *ExtraTree* (ET) fue el de mejor desempeño y por lo tanto será utilizado para la evaluación de los modelos multclasificadores determinados en el proceso que sigue a continuación.

Tabla 2. Desempeño de los clasificadores individuales.

Clasificador	ACC		AUC		Kappa	
	Entren.	Eval.	Entren.	Eval.	Entren.	Eval.
ET	0.98	0.95	0.99	0.92	0.97	0.77
NB	0.83	0.92	0.93	0.92	0.66	0.71
K	0.98	0.95	0.99	0.91	0.97	0.79
J	0.97	0.93	0.98	0.82	0.95	0.71
AD	0.94	0.92	0.97	0.93	0.88	0.71
L	0.96	0.90	0.99	0.92	0.92	0.64
MLP	0.96	0.93	0.98	0.95	0.92	0.76
SMO	0.94	0.92	0.94	0.90	0.89	0.71

Se prosiguió con la evaluación de los multclasificadores *Bagging* y *Boosting* cuyos resultados pueden ser consultados en las Tablas 3 y 4. Luego de un análisis realizado se determinó un modelo *Boosting* basado

en la red neuronal *MultiLayerPerceptron* (MLP) que supera al mejor clasificador individual evaluado y por lo tanto útil para la solución requerida. Seguidamente los multclasificadores *Stacking* y *Vote* fueron evaluados, tal como se especificó con anterioridad. Estos multclasificadores a diferencia de los anteriores no garantizan la diversidad en su aprendizaje por lo que es necesario que se garantice la diversidad en la combinación de clasificadores base que son utilizados en la obtención de los modelos multclasificadores.

Tabla 3. Desempeño de los modelos *Bagging*.

Clasificador	ACC		AUC		Kappa	
	Entren.	Eval.	Entren.	Eval.	Entren.	Eval.
ET	0.98	0.95	0.99	0.91	0.97	0.94
NB	0.85	0.92	0.96	0.96	0.71	0.95
K	0.97	0.95	0.99	0.96	0.95	0.96
J	0.97	0.92	0.99	0.97	0.95	0.96
AD	0.94	0.92	0.97	0.94	0.88	0.94
L	0.95	0.89	0.99	0.93	0.91	0.93
MLP	0.96	0.93	0.99	0.98	0.93	0.98
SMO	0.95	0.92	0.97	0.95	0.91	0.94

Tabla 4. Desempeño de los modelos *Boosting*.

Clasificador	ACC		AUC		Kappa	
	Entren.	Eval.	Entren.	Eval.	Entren.	Eval.
ET	0.98	0.95	0.99	0.97	0.97	0.98
NB	0.89	0.95	0.97	0.98	0.79	0.98
K	0.98	0.95	0.99	0.97	0.97	0.97
J	0.98	0.95	0.99	0.98	0.97	0.98
AD	0.95	0.90	0.99	0.95	0.90	0.96
L	0.97	0.90	0.99	0.95	0.94	0.96
MLP	0.98	0.96	0.99	0.99	0.97	0.98
SMO	0.96	0.90	0.99	0.96	0.92	0.96

Se procedió a determinar las combinaciones posibles a obtener con los ocho clasificadores lo cual arrojó 248 combinaciones posibles a evaluar. Las medidas de diversidad fueron calculadas y con esta información pudo ser determinado que 31 combinaciones de clasificadores eran las más diversas y por tanto las indicadas para realizar la experimentación. Estas combinaciones de clasificadores se agruparon en dos grupos, i) 20 combinaciones de clasificadores diversas en todas las medidas de diversidad calculadas y ii) 11 combinaciones diversas en cuatro de las cinco medidas de diversidad utilizadas. Información sobre los grupos conformados de combinaciones diversas a utilizar en la evaluación de los multclasificadores puede ser consultada en las Tablas 5 y 6.

Tabla 5. Combinaciones más diversas en las cinco medidas de diversidad.

R	p	D	Q	DF	Subconjunto de Clasificadores
2.00	0.47	0.06	0.94	0.03	J-MLP
2.00	0.49	0.06	0.95	0.03	ET-SMO
2.00	0.49	0.06	0.95	0.03	AD-K
2.00	0.49	0.06	0.95	0.03	ET-AD
2.00	0.49	0.06	0.95	0.03	ET-NB
2.50	0.41	0.07	0.91	0.03	J-SMO
2.50	0.41	0.07	0.91	0.03	AD-J
2.50	0.41	0.07	0.91	0.03	J-NB
2.00	0.49	0.06	0.95	0.03	K-NB

2.00	0.49	0.06	0.95	0.03	K-SMO
1.61	0.59	0.05	0.95	0.03	ET-J-SMO
1.61	0.59	0.05	0.95	0.03	ET-J-NB
1.61	0.59	0.05	0.95	0.03	ET-AD-J
1.61	0.59	0.05	0.95	0.03	AD-J-K
1.61	0.59	0.05	0.95	0.03	J-K-SMO
1.61	0.59	0.05	0.95	0.03	J-K-NB
1.64	0.57	0.05	0.95	0.04	ET-AD-J-NB
1.64	0.57	0.05	0.95	0.04	AD-J-K-NB
1.64	0.57	0.05	0.95	0.04	ET-J-NB-SMO
1.64	0.57	0.05	0.95	0.04	J-K-NB-SMO

Tabla 6. Combinaciones más diversas en cuatro de las cinco medidas de diversidad.

R	p	D	Q	DF	Subconjunto de Clasificadores
1.83	0.53	0.06	0.94	0.04	J-NB-SMO
1.83	0.53	0.06	0.94	0.04	AD-J-NB
1.53	0.57	0.06	0.95	0.04	J-L-NB
1.58	0.59	0.05	0.95	0.04	AD-J-MLP
1.67	0.61	0.05	0.94	0.04	AD-J-SMO
1.58	0.59	0.05	0.95	0.04	J-MLP-SMO
1.58	0.59	0.05	0.95	0.04	J-MLP-NB
1.56	0.61	0.05	0.95	0.04	ET-AD-J-SMO
1.56	0.61	0.05	0.95	0.04	AD-J-K-SMO
1.48	0.61	0.05	0.96	0.04	AD-J-K-NB-SMO
1.48	0.61	0.05	0.96	0.04	ET-AD-J-NB-SMO

Una vez determinada las combinaciones de clasificadores diversas se aplicó el algoritmo *Stacking* con los dos grupos de combinaciones de clasificadores. Los resultados obtenidos se muestran en las Tablas 7 y 8.

Tabla 7. Resultados del desempeño de los modelos obtenidos con el multclasificador *Stacking* sobre las combinaciones de clasificadores diversas en las cinco medidas de diversidad calculadas.

Subconjunto de Clasificadores	ACC		AUC		Kappa	
	Entren.	Eval.	Entren.	Eval.	Entren.	Eval.
J-MLP	0.93	0.92	0.93	0.90	0.87	0.71
ET-SMO	0.97	0.92	0.98	0.87	0.94	0.71
AD-K	0.97	0.90	0.97	0.85	0.95	0.64
ET-AD	0.96	0.93	0.96	0.91	0.92	0.76
ET-NB	0.97	0.95	0.97	0.88	0.95	0.79
J-SMO	0.97	0.92	0.96	0.80	0.95	0.69
AD-J	0.86	0.90	0.86	0.85	0.72	0.64
J-NB	0.97	0.92	0.97	0.90	0.94	0.71
K-NB	0.97	0.90	0.97	0.85	0.94	0.64
K-SMO	0.98	0.93	0.98	0.87	0.96	0.74
ET-J-SMO	0.93	0.90	0.95	0.85	0.86	0.61
ET-J-NB	0.98	0.89	0.98	0.84	0.97	0.60
ET-AD-J	0.97	0.92	0.97	0.86	0.95	0.69
AD--J-K	0.97	0.95	0.97	0.88	0.95	0.79
J-K-SMO	0.96	0.92	0.97	0.90	0.93	0.71
J--K-NB	0.98	0.92	0.98	0.86	0.96	0.69
ET-AD-J-NB	0.97	0.90	0.97	0.85	0.95	0.64
AD-J-K-NB	0.97	0.93	0.97	0.87	0.94	0.74
ET-J-NB-SMO	0.91	0.87	0.91	0.88	0.82	0.59
J-K-NB-SMO	0.97	0.95	0.97	0.92	0.95	0.81

Como se puede apreciar en la Tabla 7 ninguna de las combinaciones garantiza alcanzar un desempeño superior al alcanzado por mejor clasificador individual determinado con anterioridad, por tal razón no serán tenidas en cuenta para la selección del mejor modelo

multclasificador ya que en este caso el mejor resultado iguala al desempeño del mejor clasificador individual. Al realizar un análisis sobre los resultados registrados en la Tabla 8 se determinó una combinación que superó en desempeño al mejor clasificador individual, por lo que estamos en presencia del segundo modelo a utilizar como solución al problema. Esta combinación está formada por los algoritmos FDLA (AD), Jrip (J), KStar (K), NaiveBayes (NB) y SMO (SMO).

Se procedió a realizar la experimentación ahora con el algoritmo *Vote* con la utilización de los dos métodos de combinación de las salidas de los clasificadores individuales previamente especificados.

A continuación se muestran los resultados de la experimentación realizada con el multclasificador *Vote* para ambos grupos de combinaciones de clasificadores diversos registrados en las Tablas 9-12.

Tabla 8. Resultados del desempeño de los modelos obtenidos con el multclasificador *Stacking* sobre las combinaciones de clasificadores diversas en cuatro de las cinco medidas de diversidad calculadas.

Subconjunto de Clasificadores	ACC		AUC		Kappa	
	Entren.	Eval.	Entren.	Eval.	Entren.	Eval.
J-NB-SMO	0.92	0.86	0.92	0.92	0.84	0.91
AD-J-NB	0.91	0.90	0.91	0.90	0.83	0.91
J-L-NB	0.97	0.86	0.97	0.87	0.94	0.89
AD-J-MLP	0.95	0.89	0.95	0.84	0.91	0.88
AD-J-SMO	0.96	0.90	0.96	0.85	0.92	0.89
J-MLP-SMO	0.97	0.90	0.97	0.85	0.95	0.89
J-MLP-NB	0.94	0.90	0.94	0.85	0.88	0.89
ET-AD-J-SMO	0.94	0.89	0.94	0.84	0.89	0.89
AD-J-K-SMO	0.97	0.92	0.97	0.90	0.94	0.92
AD-J-K-NB-SMO	0.98	0.96	0.98	0.93	0.97	0.95
ET-AD-J-NB-SMO	0.90	0.86	0.90	0.87	0.80	0.89

Tabla 9. Resultados del desempeño de los modelos obtenidos con el multclasificador *Vote* sobre de las combinaciones de clasificadores diversas en las cinco medidas de diversidad calculadas utilizando el promedio de probabilidades como método de combinación.

Método de combinación: Promedio de probabilidades						
Subconjunto de Clasificadores	ACC		AUC		Kappa	
	Entren.	Eval.	Entren.	Eval.	Entren.	Eval.
J-MLP	0.97	0.92	0.99	0.97	0.95	0.76
ET-SMO	0.96	0.93	0.99	0.98	0.92	0.74
AD-K	0.97	0.92	0.99	0.96	0.95	0.69
ET-AD	0.98	0.95	0.99	0.96	0.97	0.79
ET-NB	0.94	0.95	0.99	0.98	0.88	0.79
J-SMO	0.94	0.92	0.99	0.92	0.89	0.71
AD-J	0.97	0.93	0.99	0.96	0.95	0.71
J-NB	0.91	0.92	0.97	0.92	0.83	0.71
K-NB	0.83	0.92	0.98	0.97	0.66	0.71
K-SMO	0.94	0.92	0.99	0.97	0.89	0.71
ET-J-SMO	0.98	0.95	0.99	0.97	0.97	0.79
ET-J-NB	0.98	0.95	0.99	0.97	0.97	0.79
ET-AD-J	0.98	0.95	0.99	0.96	0.97	0.79
AD-J-K	0.98	0.93	0.99	0.96	0.97	0.71
J-K-SMO	0.97	0.92	0.99	0.97	0.95	0.69
J-K-NB	0.97	0.92	0.99	0.97	0.95	0.69
ET-AD-J-NB	0.98	0.95	0.99	0.97	0.97	0.79
AD-J-K-NB	0.97	0.92	0.99	0.97	0.95	0.69
ET-J-NB-SMO	0.95	0.93	0.99	0.97	0.91	0.76
J-K-NB-SMO	0.95	0.93	0.99	0.97	0.91	0.76

Tabla 10. Resultados del desempeño de los modelos obtenidos con el multclasificador Vote sobre las combinaciones de clasificadores diversas en las cinco medidas de diversidad calculadas utilizando el voto mayoritario como método de combinación.

Método de combinación: Voto mayoritario						
Subconjunto de Clasificadores	ACC		AUC		Kappa	
	Entren.	Eval.	Entren.	Eval.	Entren.	Eval.
J-MLP	0.97	0.92	0.97	0.81	0.94	0.66
ET-SMO	0.96	0.92	0.96	0.86	0.93	0.69
AD-K	0.96	0.92	0.96	0.86	0.93	0.69
ET-AD	0.96	0.92	0.96	0.86	0.93	0.69
ET-NB	0.92	0.92	0.92	0.98	0.84	0.79
J-SMO	0.96	0.90	0.96	0.80	0.92	0.61
AD-J	0.96	0.90	0.96	0.80	0.92	0.61
J-NB	0.91	0.90	0.91	0.80	0.83	0.61
K-NB	0.92	0.92	0.92	0.86	0.84	0.69
K-SMO	0.96	0.92	0.96	0.86	0.93	0.69
ET-J-SMO	0.99	0.95	0.99	0.88	0.98	0.79
ET-J-NB	0.99	0.95	0.99	0.88	0.98	0.79
ET-AD-J	0.99	0.95	0.99	0.88	0.98	0.79
AD-J-K	0.98	0.95	0.98	0.88	0.97	0.79
J-K-SMO	0.98	0.95	0.98	0.88	0.97	0.79
J-K-NB	0.98	0.95	0.98	0.88	0.97	0.79
ET-AD-J-NB	0.97	0.93	0.97	0.87	0.94	0.74
AD-J-K-NB	0.97	0.93	0.97	0.87	0.94	0.74
ET-J-NB-SMO	0.97	0.93	0.97	0.87	0.94	0.74
J-K-NB-SMO	0.97	0.93	0.97	0.87	0.94	0.74

Tabla 11. Resultados del desempeño de los modelos obtenidos con el multclasificador Vote sobre las combinaciones de clasificadores diversas en cuatro de las cinco medidas de diversidad calculadas utilizando el promedio de probabilidades como método de combinación.

Método de combinación: Promedio de probabilidades						
Subconjunto de Clasificadores	ACC		AUC		Kappa	
	Entren.	Eval.	Entren.	Eval.	Entren.	Eval.
J-NB-SMO	0.95	0.93	0.98	0.92	0.91	0.76
AD-J-NB	0.94	0.93	0.98	0.97	0.89	0.76
J-L-NB	0.96	0.92	0.99	0.97	0.92	0.69
AD-J-MLP	0.97	0.93	0.99	0.97	0.95	0.76
AD-J-SMO	0.94	0.92	0.99	0.96	0.89	0.71
J-MLP-SMO	0.96	0.93	0.99	0.97	0.92	0.76
J-MLP-NB	0.96	0.93	0.99	0.97	0.93	0.76
ET-AD-J-SMO	0.98	0.95	0.99	0.97	0.97	0.98
AD-J-K-SMO	0.97	0.92	0.99	0.97	0.95	0.69
AD-J-K-NB-SMO	0.95	0.93	0.99	0.97	0.91	0.76
ET-AD-J-NB-SMO	0.95	0.93	0.99	0.97	0.91	0.76

Tabla 12. Resultados del desempeño de los modelos obtenidos con el multclasificador Vote sobre las combinaciones de clasificadores diversas en cuatro de las cinco medidas de diversidad calculadas utilizando el voto mayoritario como método de combinación.

Método de combinación: Voto mayoritario						
Subconjunto de Clasificadores	ACC		AUC		Kappa	
	Entren.	Eval.	Entren.	Eval.	Entren.	Eval.
J-NB-SMO	0.95	0.93	0.95	0.91	0.91	0.76
AD-J-NB	0.95	0.93	0.95	0.91	0.91	0.76
J-L-NB	0.96	0.92	0.96	0.96	0.92	0.69
AD-J-MLP	0.96	0.93	0.96	0.91	0.92	0.76
AD-J-SMO	0.94	0.92	0.94	0.90	0.89	0.71
J-MLP-SMO	0.96	0.93	0.96	0.91	0.92	0.76
J-MLP-NB	0.96	0.93	0.96	0.91	0.93	0.76

ET-AD-J-SMO	0.96	0.92	0.96	0.86	0.93	0.69
AD-J-K-SMO	0.96	0.93	0.96	0.86	0.93	0.69
AD-J-K-NB-SMO	0.95	0.93	0.95	0.91	0.91	0.76
ET-AD-J-NB-SMO	0.95	0.93	0.95	0.76	0.91	0.91

Como resultado del análisis de la información registrada en las Tablas 9-12 no se identificó ninguna combinación que superara al desempeño del mejor clasificador individual, por tal razón no son soluciones que serán tenidas en cuenta para la solución del problema.

Selección de los modelos aptos para la solución del problema

Al realizar un análisis de los resultados obtenidos a partir de los multclasificadores se identificaron dos modelos multclasificadores que mostraron un desempeño en el conjunto de evaluación externa superior al desempeño logrado por el algoritmo *ExtraTree* (ver Tabla 13). A partir del análisis realizado con los multclasificadores de mejor desempeño se pudo determinar que, los modelos no presentan diferencias significativas en su desempeño, ya que poseen el mismo valor de la exactitud (ACC) y una diferencia no significativa en los parámetros referentes al área bajo la curva ROC y el estadístico Kappa.

En la modelación clásica estadística donde normalmente se emplean modelos lineales, la capacidad generalizadora de estos es seriamente afectada por la relación entre instancias y variables descriptivas. Sin embargo, el empleo de algoritmos de aprendizaje automático tiene la ventaja, además de explorar relaciones no lineales, de ser mucho menos sensible a esta relación. El hecho de que todos los modelos base y multclasificadores generados muestren un desempeño muy similar en ambos conjuntos (entrenamiento y evaluación externa) demuestra que efectivamente, el empleo de estos algoritmos permite generar modelos sin sobreajuste, independientemente de la relación entre casos y variables descriptivas (ver Tablas 2 – 4 y 7 -13).

Así mismo, si consideramos los requerimientos según la Organización para la Cooperación Económica y el Desarrollo (OECD) para el uso de modelos QSAR como herramientas alternativas para la evaluación del peligro potencial de sustancias y compuestos químico^{48,49}, podemos afirmar que los modelos generados cumplen con la mayoría de estos requerimientos. La obtención de los modelos propuestos (Tabla 13) está descrita detalladamente por lo que cumple con el criterio de ser un algoritmo no ambiguo. Así mismo están destinados a una aplicación bien específica: la evaluación del potencial neurotóxico de LI a través de la estimación de su capacidad inhibitoria sobre la enzima AChE. Por tanto, cumple con los requerimientos de tener un objetivo bien definido y un dominio de aplicación establecido. Aunque este último no es determinado explícitamente, la aplicación de cortes de similitud arbitrarios usando como espacio de referencia el vector reducido de 50 fragmentos generados con ISIDA, puede ayudar a decidir que nuevos LI están o no dentro del dominio de aplicación de los modelos propuestos.

En las Tablas 2 – 4 y 7 -13 es claro que los modelos descritos en cada una de las tablas muestran ex-

celentes valores de exactitud y predictividad, por lo que cumple con uno de los principales requerimientos encaminados a asegurar la capacidad generalizadora de los modelos QSAR. Finalmente, nuestros modelos (sobre todo los multclasificadores) al ser “cajas negras” carecen que la interpretación mecanística requerida en el quinto principio de la OECD. No obstante, este principio no es de obligatorio cumplimiento, de acuerdo con la OECD.

Todo lo anterior sugiere que ambos modelos son idóneos en la solución del problema, por lo que constituyen las propuestas a utilizar para la evaluación del perfil neurotóxico de los LI (seguros o tóxicos) basado en inhibición de la enzima AChE.

Tabla 13. Resultados de los mejores multclasificadores y el mejor clasificador individual.

	ACC		AUC		Kappa	
	Entren.	Eval.	Entren.	Eval.	Entren.	Eval.
Mejor clasificador individual (ET)	0.98	0.95	0.99	0.92	0.97	0.77
Multclasificador AdaBoostM1	0.98	0.96	0.99	0.99	0.97	0.98
Multclasificador Stacking	0.98	0.96	0.98	0.93	0.97	0.95

CONCLUSIONES

En el trabajo fueron obtenidos dos modelos, basados en multclasificadores, que permitieron predecir el potencial neurotóxico de nuevas moléculas de líquidos iónicos basado en el grado de inhibición de la enzima Acetilcolinesterasa. Para la obtención del primer modelo se empleó el algoritmo *AdaBoostM.1* usando como clasificador base una red neuronal *Multilayer-Perceptron*. Para el segundo modelo se empleó el algoritmo *Stacking* y la combinación de cinco clasificadores base (*FDLA*, *Jrip*, *Kstar*, *NaiveBayes* y *SMO*). Esta combinación de clasificadores fue seleccionada mediante la aplicación del cálculo de medidas de diversidad.

Del análisis del desempeño de los modelos obtenidos es posible esperar que al menos el 96% de las instancias que conforman un nuevo conjunto de líquidos iónicos será clasificado correctamente. Dichos modelos constituyen herramientas útiles para la predicción temprana del perfil neurotóxico de nuevos líquidos iónicos.

AGRADECIMIENTOS

Bolsa de Postdoctorado [SFRH/BPD/90673/2012] financiada por la FCT - Fundación para la Ciencia y la Tecnología, Portugal, cofinanciada por el Fondo Social Europeo.

BIBLIOGRAFIA

1. Kokorin, A. Editor. Ionic Liquids: Theory, Properties, New Approaches. Rijeka, Croatia: *InTech*. 2011.

2. Holbrey, J.D.; Seddon, K.R. Ionic liquids. *Clean. Prod. Proces.* **1999**, *1*, 223-237.
3. Welton, T. Room-temperature ionic liquids. Solvents for synthesis and catalysis. *Chem. Rev.* **1999**, *99*, 2071-2084.
4. Wasserscheid, P.; Welton, T. Ionic liquids in synthesis. *Weinheim: Wiley-VCH*. **2003**.
5. Rogers, R.D.; Seddon, K.R. Ionic liquids as green solvents: Progress and prospects. *American Chemical Society (ACS)*. **2003**, 620
6. Ranke, J.; Stolte, S.; Stormann, R.; Arning, J.; Jastorff, B. Design of sustainable chemical products - the example of ionic liquids. *Chem. Rev.* **2007**, *107*, 2183-2206.
7. Pham, T.P.; Cho, C.W.; Yun, Y.S. Environmental fate and toxicity of ionic liquids: a review. *Water. Res.* **2010**, *44*, 352-372.
8. Frade, R.F.; Afonso, C.A. Impact of ionic liquids in environment and humans: an overview. *Hum. Exp. Toxicol.* **2010**, *29*, 1038-1054.
9. Zhao, D.; Liao, Y.; Zhang, Z. Toxicity of ionic liquids. *Clean.* **2007**, *35*, 42-48.
10. Stock, F.; Hoffmann, J.; Ranke, J.; Stormann, R.; Ondruschka, B.; Jastorff, B. Effects of ionic liquids on the acetylcholinesterase -A structure-activity relationship consideration. *Green Chem.* **2004**, *6*.
11. Arning, J.; Stolte, S.; Boschen, A.; Stock, F.; Pitner, W.; Welz, U.; et al. Qualitative and quantitative structure activity relationship for the inhibitory effects of cationic head groups, functionalised side chains and anions of ionic liquids on acetylcholinesterase. *Green Chem.* **2008**, *10*, 11.
12. García, A.; Tojo, E.; Tojo, J.; Teijeira, M.; Rodríguez, F.J.; González, M.P.; et al. Cytotoxicity of selected imidazolium-derived ionic liquids in the human Caco-2 cell line. Sub-structural toxicological interpretation through a QSAR study. *Green Chem.* **2008**, *10*, 8.
13. Torrecilla, J.; Palomar, J.; Lemus, J.; Rodríguez, F. A quantum-chemical-based guide to analyze/quantify the cytotoxicity of ionic liquids. *Green Chem.* **2010**, *12*, 11.
14. Couling, D.; Bernot, R.J.; Docherty, K.M.; Dixon, J.K.; Maginn, E.J. Assessing the factors responsible for ionic liquid toxicity to aquatic organisms via quantitative structure-property relationship modeling. *Green Chem.* **2006**, *8*.
15. Luis P.; Ortiz, I.; Aldaco, R.; Irabien, A. A novel group contribution method in the development of a QSAR for predicting the toxicity (*Vibrio fischeri* EC50) of ionic liquids. *Ecotox. Environ. Safe.* **2007**, *67*, 6.
16. Lacrama, A.; Putz, M.; Ostafe, V. A Spectral-SAR model for the anionic-cationic interaction in ionic liquids: Application to *vibrio fischeri* ecotoxicity. *Int. J. Mol. Sci.* **2007**, *8*, 21.
17. Putz, M.; Lacrama, A.; Ostafe, V. Spectral SAR ecotoxicology of ionic liquids: The daphnia magna case. *Int. J. Ecology.* **2007**, *7*.
18. Mohammad, H.; Fatemi, A.; Izadiyan, P. Cytotoxicity estimation of ionic liquids based on their

- effective structural features. *Chemosphere*. **2011**, *84*, 10.
19. Luis, P.; Garea, A.; Irabien, A. Quantitative structure – activity relationships (QSARs) to estimate ionic liquids ecotoxicity EC50 (*Vibrio fischeri*). *J. Mol. Liq.* **2010**, *152*, 5.
 20. Hossain, M.; Belhaouari, B.; El-Harbawi, M.; Nusaibah, A.; Abdul, M.I.; Hefter, G.; et al. Development of a novel mathematical model using a group contribution method for prediction of ionic liquid toxicities. *Chemosphere*. **2011**, *85*, 4.
 21. Mohammad, H.; Fatemi, A.; Izadiyan, P. In silico cytotoxicity estimation of ionic liquids based on their two- and three-dimensional structural descriptors. *Monatsh Chem.* **2011**, *142*, 8.
 22. Frade, R.F.; Afonso, C.A. Impact of ionic liquids in environment and humans: an overview. *Human and Experimental Toxicology (HET)*. **2010**, *29*, 1038–1054.
 23. Störmann, R.; Arning, J.; Stolte, S.; Matzke, M.; Jastorff, B.; Thöming, J. UFT Merck Ionic Liquids Biological Effects Database **2014** [cited 2014]. Available from: <http://www.il-eco.uft.uni-bremen.de>.
 24. Tropsha, A.; Fourche, D.; Muratov, E. Trust, but verify: On importance of chemical structure curation in cheminformatic and QSAR modeling research. *J. Chem. Inf. Model.* **2010**, *50*, 1189–1204.
 25. ChemAxon. JChem for Excel. version 5.10.2.725 ed**2012**.
 26. ChemAxon. Standardizer. version 5.10.2 ed**2012**.
 27. Varnek, A.; Fourche, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P. et al. ISIDA-platform for virtual screening based on fragment and pharmacophore descriptors. *Curr. Comput. Aided Drug Des.* **2008**, *4*.
 28. Ruggiu, F.; Marcou, G.; Solov'ev, V.; Horvath, D.; Varnek, A. ISIDA Fragmentor 2013. Laboratoire de Chémo-informatique, Université de Strasbourg, France.2013. p. Tools for the calculation of molecular fragment count descriptors from a Structure-Data File (SDF).
 29. Cruz-Monteagudo, M.; Ancede-Gallardo, E.; Jorge, M.; Dias, S.; Cordeiro, M.N. Chemoinformatics profiling of ionic liquids--automatic and chemically interpretable cytotoxicity profiling, virtual screening, and cytotoxicophore identification. *Toxicol Sci.* **2013**, *136*, 548-565.
 30. WEKA. Waikato Environment for Knowledge Analysis (WEKA). 3.7.11 ed. New Zealand: University of Waikato; 2014.
 31. StatSoft. STATISTICA. version 8.0 ed2007. p. (data analysis software system).
 32. Guo, H.; Viktor, H.L. Learning from Imbalanced Data Sets with Boosting and Data Generation: The DataBoost-IM Approach. *Sigkdd Explorations*. **2004**, *6*, 30.
 33. Aguiar, V.; Gestal, M.; Cruz-Monteagudo, M.; Rabuñal, J.R.; Dorado, J.; Munteanu, C.R. Evolutionary computation and QSAR research. *Curr. Comput. Aided Drug Des.* **2013**, *9*, 19.
 34. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance and Min-Redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*.
 35. Ramentol, E.; Caballero, Y.; Bello, R.; Herrera, F. SMOTE-RS B*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory. *Knowl. Inf. Syst.* **2012**, *33*, 20.
 36. Chawla, N.; Bowyer, K.; Hall, L.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*.
 37. Hall, M.; Frank, H.; Holmes, H.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA Data Mining Software: An Update. *ACM SIGKDD explorations newsletter*. **2009**, *11*.
 38. Breiman, L.; Bagging predictors. *Machine Learning*. **1996**, *24*, 17.
 39. Freund, Y.; SCHAPIRE RE. A Short Introduction to Boosting. *Trans. Jpn. Soc. Artif. Intell.* **1999**, *14*, 9.
 40. Wolper, D.H. Stacked Generalization. *Neural Networks*. **1992**; *5*, 18.
 41. Casas, G.; Morales, J.L.; Denoda, L.; Cabrera, L.; Arguelles, L.; Gonzalez, E. High cardiovascular risk models for a city in the central region of Cuba. *Age*. **2013**, *18*.
 42. Kuncheva, L.I. Combining Pattern Classifiers: Methods and Algorithms: *John Wiley & Sons*. **2004**, *964*.
 43. Bonet, I.; Franco-Montero, P.; Rivero, V.; Tejeira, M.; Borges, F.; Uriarte, E., et al. Classifier ensemble based on feature selection and diversity measures for predicting de affinity of A2B Adenosine Receptor Antagonist. *J. Chem. Inf. Model.* **2013**, *53*, 16.
 44. García, S.; Fernández, A.; Luengo, L.; Herrera, F. A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability. *Soft Comput.* **2009**, *13*.
 45. Witten, I.H.; Frank, E.; Data Mining. A practical machine learning tools and techniques. Segunda edición ed2005 **2005**, *558*.
 46. Billsus, D.; Pazzani, M.J. Learning Collaborative Information Filters. **1998**.
 47. Richert, W.; Coelho, L.P. Building machine learning systems with Python. *Packt Publishing Ltd*. **2013**.
 48. Zeeman, M.; Auer, C.M.; Clements, R.G.; Nabolholz, J.V.; Boethling, R.S. U.S. EPA regulatory perspectives on the use of QSAR for new and existing chemical evaluations. *SAR QSAR Environ. Res.* **1995**, *3*, 179-201.
 49. Worth, A.P.; Van Leeuwen, C.J.; Hartung ,T. The prospects for using (Q)SARs in a changing political environment--high expectations and a key role for the European Commission's joint research centre. *SAR QSAR Environ. Res.* **2004**, *15*, 331-343.