УДК: 004.032.26+620.26

Ерёмина В.С., Михайлова П.Г.

РАЗРАБОТКА МОДЕЛЕЙ ПРОГНОЗИРОВАНИЯ ПОКАЗАТЕЛЕЙ ОПАСНОСТИ ХИМИЧЕСКОЙ ПРОДУКЦИИ В УСЛОВИЯХ НЕОПРЕДЕЛЕННОСТИ

Ерёмина Виктория Сергеевна, студентка 2 курса магистратуры факультета информационных технологий и управления, e-mail: eremvik@gmail.com;
Михайлова Павла Геннадьевна, к.т.н., доцент, доцент кафедры компьютерно-интегрированных систем в химической технологии;
Российский химико-технологический университет им. Д.И. Менделеева, Москва, Россия
125480, Москва, ул. Героев Панфиловцев, д. 20

В данной работе рассмотрен процесс разработки моделей прогнозирования показателей опасности химической продукции с использованием методологии QSAR, а также нейронных сетей в условиях неопределенности, т.е. при отсутствии или противоречивости исходных данных. Приведены результаты расчетов показателей острой токсичности и экотоксичности для нитрозобензола и циклогексилгидроксиамина.

Ключевые слова: *нейронные сети, показатели опасности, химическая продукция, КССА.*

DEVELOPMENT OF MODELS OF PREDICTION OF HAZARD OF CHEMICAL PRODUCTS IN UNCERTAINTY OF THE ORIGINAL DATA

Eremina V.S., Mikhaylova P.G.
D. Mendeleev University of Chemical Technology of Russia, Moscow, Russia

In this work, the development of models of prediction of hazard of chemical products using the methodology of QSAR and neural networks in uncertainty conditions, i.e., in the absence or inconsistency of the original data. The results of calculations of indicators of acute toxicity and ecotoxicity for nitrosobenzene and cyclohexylhydroxylamine.

Keywords: *neural networks, indicators of danger, chemical products, QSAR.*

Жизнь современного человека невозможно представить без ежедневного использования различных химических веществ и их смесей. Каждый год во всем мире производят огромное количество разнообразной химической продукции, а так же получают и синтезируют около 1000 новых соединений. Но их производство, хранение, использование и утилизация влекут за собой целый ряд опасностей как для человека, так и для окружающей среды. Чтобы предотвратить возможные негативные последствия, а также снизить риск их возникновения при обращении продукции в России, а также в странах Европейского Союза был разработан ряд нормативных документов, регламентирующих безопасное обращение химической продукции. Следуя им, производители обязаны наносить на упаковку предупредительную маркировку, информирующую потребителя о характеристиках опасности и мерах предосторожности при использовании химической продукции.

Ключевым моментом для разработки предупредительной маркировки является классификация по видам опасности. Выделяют следующие виды опасности: по физико-химическим показателям, токсичности для человека и экотоксичности. Для проведения классификации по острой токсичности для человека и экотоксичности необходимо знать значения показателей опасности, таких как, *LD*50 (*letal dose*) – среднесмертельная доза вещества при введении в желудок и при нанесении его на кожу; *LC*50 (*letal concentration*) – среднесмертельная концентрация вещества в воздухе при ингаляционном воздействии; *CL*50 – среднесмертельная концентрация вещества в воде для определения острой токсичности для водных организмов (рыб и водорослей) [1].

Есть много баз данных (БД), содержащих информацию по данным видам опасности. Показатели опасности определяют экспериментальным путем, но этот процесс достаточно продолжительный и дорогостоящий, а проведение экспериментов на животных (они необходимы для определения показателей острой токсичности для человека и экотоксичности) противоречит этическим нормам, к тому же создать одинаковые условия для проведения таких испытаний достаточно сложно. Разные БД могут содержать противоречивую информацию для одного и того же вещества или такая информация может отсутствовать вовсе. Такие условия можно считать неопределенными. Поэтому разработка моделей для прогнозирования показателей опасности химической продукции является актуальной в настоящее время.

Рассмотрим некоторые методы прогнозирования. В случае, когда экспериментальные данные отсутствуют, можно произвести прогнозирование показателей опасности по методологии *QSAR* (*Quantitative Structure-Activity Relationship – количественное соотношение структура-активность, КССА*), которая основана на предположении о существовании связи строения веществ с их свойствами (активностью). В особенности эта методология широко используется

для органических соединений. Разработчики Европейского Химического Агентства (*European Chemicals Agency, ECHA*) создали свободно распространяемую программу *QSAR Toolbox* [2], которая позволяет заполнять пробелы в данных по токсичности и экотоксичности. Она включает в себя информацию из многих БД и производит расчет путем нахождения и группировки веществ-аналогов для исследуемого вещества, основываясь на его структурной формуле. Затем строится зависимость необходимого показателя опасности веществ-аналогов от какого-либо физико-химического свойства и проводится прогнозирование методами трендового анализа или «чтения по диагонали». Такой способ прогнозирования является безопасным, а также быстрым, и занимает от 2 до 5 часов, время зависит от сложности структуры соединения. В то время как при определении показателей токсичности экспериментально испытуемые наблюдаются от 24 часов до нескольких суток.

Еще один метод прогнозирования – использование нейросетевого анализа данных. В данной работе для прогнозирования показателей острой токсичности химической продукции для человека и окружающей среды предлагается использовать нейронные сети (НС) прямого распространения. Для прогнозирования выбраны следующие 4 показателя: среднесмертельные дозы при введении в желудок *LD*50(в/ж) и при нанесении на кожу *LD*50(н/к), среднесмертельная концентрация при вдыхании *LC*50(инг), среднесмертельная концентрация вещества в воде для водных организмов *CL*50. Для этих показателей в соответствии с рекомендациями [1] были определены влияющие на них физико-химические свойства веществ:

$$LD50(\text{в/ж}) = f_1(M, \rho, T_{\text{кип}}), \quad (1)$$

$$LD50(\text{н/к}) = f_2(M, \rho, T_{\text{кип}}, \lg K_{\text{ow}}), \quad (2)$$

$$LC50(\text{инг}) = f_3(M, \rho, P_{\text{нас.пара}}), \quad (3)$$

$$CL50 = f_4(M, \lg K_{\text{ow}}, S, K_{\text{г}}), \quad (4)$$

где M – молярная масса вещества, г/моль;
ρ – плотность при 20°С, г/мл;
$T_{\text{кип}}$ – температура кипения при 760 мм.рт.ст, °С;
$\lg K_{\text{ow}}$ – десятичный логарифм от коэффициента разделения октанол-вода;
$P_{\text{нас.пара}}$ – давление насыщенного пара при 25°С, мм.рт.ст;
S – растворимость при 20°С, мг/л;
$K_{\text{г}}$ – константа Генри при 25°С, атм*м^3/моль.

Таким образом, показатели токсичности и экотоксичности являются выходами НС, а физико-химические свойства – входами. Для моделей (1) и (3) НС имеют следующие структуры: 3 входа, 1 выход и 1 скрытый слой (рис.1); для моделей (2) и (4): 4 входа, 1 выход и 1 скрытый слой (рис.2). На рис. 1 и 2 использованы обозначения: x_1-x_4 – входы НС, w_{nk} – весовые коэффициенты связей между входами НС (n = 1-4) и нейронами скрытого слоя (k = 1-j), v_{k1} – весовые коэффициенты связей между нейронами скрытого и выходного слоя, y – нейрон выходного слоя. Количество нейронов в скрытом слое подбирается отдельно для каждой сети.

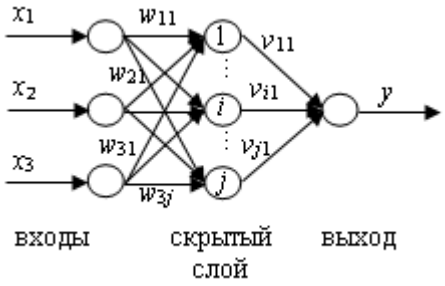

Рис.1. Модель нейронной сети с тремя входами, одним скрытым слоем и одним выходом

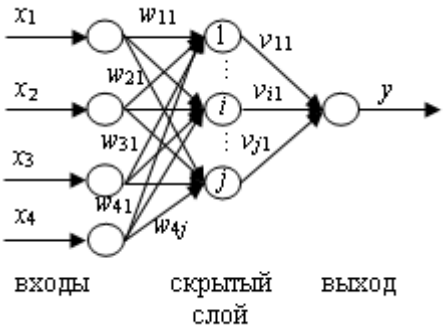

Рис.2. Модель нейронной сети с четырьмя входами, одним скрытым слоем и одним выходом

Для реализации предложенных моделей в данной работе была использована программа российской компании BaseGroup Labs – *Deductor Studio Academic* 5.3.0.88 [3]. Эта версия программы является образовательной, она бесплатная и общедоступная.

По данным о веществах-аналогах сформируется общая выборка, в которую входят выбранные показатели опасности и необходимые значения физико-химических свойств. Далее, общая выборка разбивается на две: обучающую и тестовую. В обучающую входят 90% данных из общей выборки, а в тестовую – оставшиеся 10%. На следующем этапе проводится обучение НС. Чтобы подобрать наилучшую сеть, изменяется количество нейронов в скрытом слое, затем функция активации, ее крутизна, метод и скорость обучения. На основе полученных значений о количестве распознанных примеров в выборках, а также средней ошибки обучения, варьируются параметры настройки сети до тех пор, пока не будут достигнуты лучшие показатели. Когда сеть выбрана, с помощью функции «*что-если*» можно ввести необходимые входные значения, и программа выдаст рассчитанное значение выхода, т.е. искомый показатель опасности.

В качестве примеров для прогнозирования показателей опасности были выбраны следующие вещества: нитрозобензол (CAS 586-96-9),

циклогексилгидроксиламин (CAS 2211-64-5), имеющие схожее строение: наличие шестичленного цикла, а также одного атома азота, связанного с кислородом.

Для них была составлена выборка веществ-аналогов, основываясь на данных, полученных при расчете с помощью программы *QSARToolbox*. Затем недостающая информация о веществах была получена из БД по свойствам химических веществ: *TOXNET* [4] и БД по паспортам безопасности химической продукции: *MOLBASE* [5], *Sigma-Aldrich* [6]. Общий объем выборки по всем свойствам составил около 140 веществ.

Те же самые показатели были рассчитаны в программе *QSARToolbox*. Результаты расчетов представлены в таблице 1.

Таблица 1. Результаты прогнозирования показателей опасности

Вещество	Показатель опасности		Программа		
			QSARToolbox		*Deductor*
			Метод расчета		
			Organic Functional Group	Structure Similar	Нейронные сети прямого распространения
Нитрозобензол	Для человека	*LD*50(в/ж), мг/кг (крысы)	373	638	479
		*LD*50(н/к), мг/кг (кролики)	–	1650	1571
		*LC*50(инг), мг/л (крысы, 4 ч)	–	0,34	0,34
	Для окружающей среды	*CL*50, мг/л (96 ч)	–	15,2	30
Циклогексилгидроксиламин	Для человека	*LD*50(в/ж), мг/кг (крысы)	13,3	1100	1069
		*LD*50(н/к), мг/кг (кролики)	–	939	1045
		*LC*50(инг), мг/л (крысы, 4 ч)	–	35,3	31,2
	Для окружающей среды	*CL*50, мг/л (96 ч)	–	35,2	32,1

По результатам расчетов можно сделать вывод о том, что используемые методы являются достаточно точными, так как значения, полученные разными способами, практически совпадают.

Стоит отметить, что значения среднесметрельной дозы при введении в желудок для циклогексилгидроксиламина, полученные при расчете по программе *QSARToolbox*, отличаются почти в 10 раз (находятся в пределах от 13,3 мг/кг до 1100 мг/кг). Это может быть связано с тем, что по разным методам расчета этой программы (Organic Functional Group – органические функциональные группы и Structure Similar – структурная схожесть) были найдены разные вещества-аналоги, а для прогнозирования с помощью НС вещества-аналоги были объединены в одну выборку (значение 1069 мг/кг).

Метод прогнозирования с использованием нейросетевого моделирования является более предпочтительным, так как НС может аппроксимировать произвольные функции нескольких аргументов, определенных на некотором множестве, а в методах прогнозирования, используемых в программе *QSARToolbox*, заложены линейные зависимости функции одной переменной.

Список литературы

1. Савицкая Т.В., Егоров А.Ф., Михайлова П.Г. Классификация химических опасностей: методы, критерии, показатели: учеб.пособие. М: РХТУ им. Д.И.Менделеева, 2010. 148 с.

2. QSAR Toolbox 3.4 // Organisation for Economic Co-operation and Development & European Chemicals Agency. 2012 [Электронный ресурс]. Режим доступа: https://www.qsartoolbox.org/download (дата обращения: 15.04.2017).

3. Deductor // BaseGroup Labs ООО «Аналитические технологии». 2017 [Электронный ресурс]. Режим доступа: https://basegroup.ru/deductor/description (дата обращения: 25.04.2017).

4. TOXNET // US National Library of Medicine. 1993 [Электронный ресурс]. Режим доступа: https://toxnet.nlm.nih.gov/ (дата обращения: 17.04.2017).

5. MOLBASE // MOLBASE (Shanghai) Biotechnology Co.,Ltd. 2013 [Электронный ресурс]. Режим доступа: http://www.molbase.com/ (дата обращения: 17.04.2017).

6. Sigma-Aldrich // Merck Group. 2015 [Электронный ресурс]. Режим доступа: http://www.sigmaaldrich.com (дата обращения: 17.04.2017).