УДК 004.89

Самигулина Г.А., Самигулина З.И.
Институт информационных и вычислительных технологий, Казахстан, Алматы

ПРИМЕНЕНИЕ СОВРЕМЕННЫХ МЕТОДОВ DATA MINING ДЛЯ ПРОГНОЗИРОВАНИЯ ЗАВИСИМОСТИ «СТРУКТУРА/СВОЙСТВО» ХИМИЧЕСКИХ СОЕДИНЕНИЙ СУЛЬФАНИЛАМИДОВ

Аннотация. В настоящее время актуальна разработка современных интеллектуальных технологий прогнозирования новых лекарственных препаратов с заданными фармакологическими свойствами на базе методов искусственного интеллекта и статистического анализа данных. Статья посвящена исследованиям в области компьютерного молекулярного дизайна новых лекарственных препаратов сульфаниламидов с применением современного программного обеспечения для интеллектуального анализа данных Rapid Miner. Разработана база данных дескрипторов сульфаниламидов на основе крупнейшего мирового ресурса химической информации Mol-Instincts. Представлена графическая модель предварительной обработки дескрипторов сульфаниламидов в среде Rapid Miner с помощью алгоритма Random Forest и метода главных компонент. Получены результаты численного моделирования, осуществлена визуализация данных в 2D и 3D форме.

Ключевые слова: data mining, дескрипторы сульфаниламидов

Введение

В настоящее время открыто огромное количество структур химических соединений, в связи с чем остро стоит проблема создания современных интеллектуальных технологий, позволяющих обрабатывать большой массив химической информации. Хорошо зарекомендовали себя методы искусственного интеллекта (ИИ) и статистического анализа данных для решения задач прогнозирования зависимости «структура-активность» химических соединений (Quantitative Structure-Activity Relationship, QSAR). К ним относятся: нейронные сети (НС) [1], генетические алгоритмы (ГА) [2], искусственные иммунные системы (ИИС) [3,4] и т.д. Так же широкое применение получили: метод опорных векторов [5], метод «ближайшего соседа» [6], метод главных компонент [7], алгоритм на основе ансамбля деревьев решений [8] и т.д. Данные алгоритмы интеллектуального анализа данных успешно применяются для решения задач предварительной обработки данных, распознавания образов и прогнозирования при синтезе новых лекарственных препаратов.

Алгоритмы Data Mining для прогнозирования зависимости «структура/свойство» химических соединений сульфаниламидов

Рассмотрим основные принципы интеллектуального анализа баз данных химической информации (Рисунок 1). В настоящее время результат решения задачи распознавания образов различными алгоритмами во многом зависит от исходного набора данных. Структуры химических соединений описываются с помощью различных дескрипторов (структурных, топологических, геометрических, квантово-химических и т.д.).

В результате получаются базы данных огромного размера, обработка которых является трудоемкой задачей. В связи с этим актуально осуществление предварительной обработки данных. В качестве алгоритмов для предварительной обработки данных хорошо зарекомендовали себя метод главных компонент и алгоритм Random Forest (Рисунок 1).

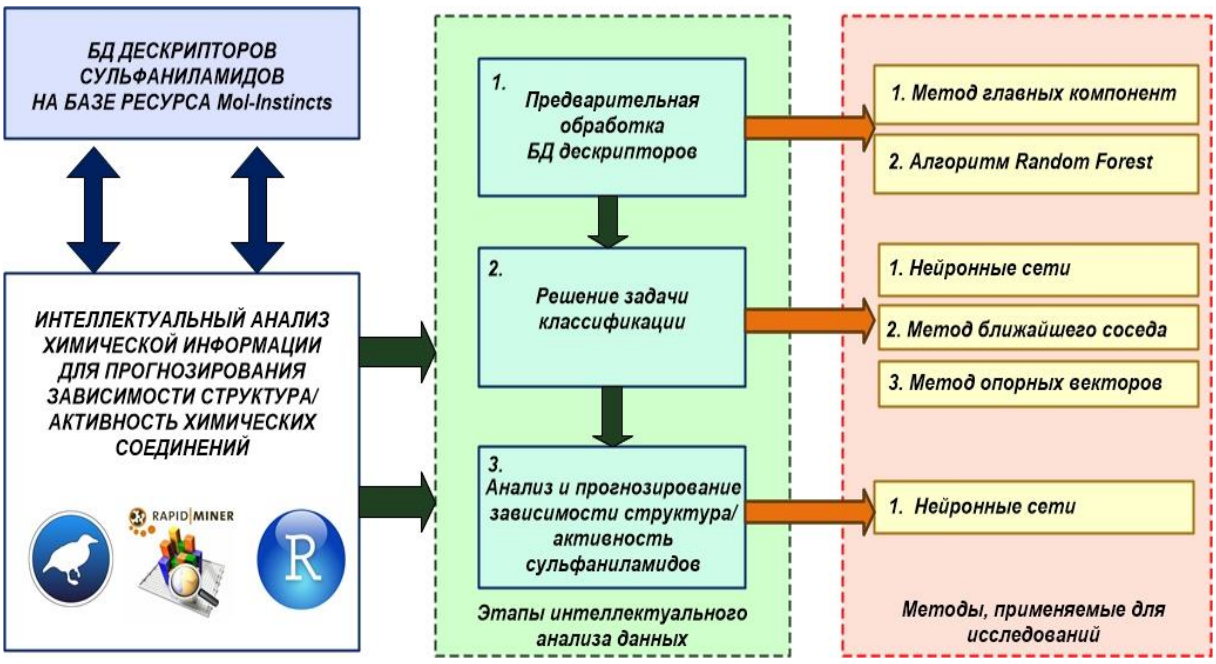

Рисунок 1 – Этапы интеллектуального анализа химической информации для прогнозирования зависимости «структура/свойство» сульфаниламидов

Метод главных компонент позволяет уменьшить размерность данных с потерей минимального количества информации, а одним из достоинств является способность выявления скрытых (латентных) свойств признаков. Другой алгоритм на основе ансамбля деревьев решений Random Forest, разработанный Л. Брейманом, представляет собой комбинацию деревьев решений, которые являются числовым параметром метода, таким образом, что каждое дерево зависит от значения случайного вектора независимой выборки с таким же распределением, как и у всех деревьев в лесу [9]. Брейманом предложены четыре меры информативности дескрипторов, которые позволяют ранжировать переменные по степени значимости. Достоинствами данного метода являются возможность обработки данных с большим числом признаков и классов, а так же нечувствительность к масштабированию.

После этапа предварительной обработки базы данных химической информации осуществляется решение задачи классификации (Рисунок 1). Для этих целей интересны исследования алгоритмов на основе нейронных сетей, метода опорных векторов и метода «ближайшего соседа». Нейронные сети имитируют способность биологических нервных систем обучаться и исправлять ошибки. Данный алгоритм широко используется в хемоинформатике для исследования количественного соотношения «структура/свойство» химических соединений.

Метод «ближайшего соседа», (nearest neighbour) относится к классу методов, работа которых основывается на хранении данных в памяти для сравнения с новыми элементами [10]. При появлении новой записи для прогнозирования находятся отклонения между этой записью и подобными наборами данных, и наиболее подобная (или ближний сосед) идентифицируется. Метод является простым в использовании и хорошо подходит для решения задач небольшой размерности.

Метод опорных векторов относится к классу граничных методов для решения задач бинарной классификации, где определение классов осуществляется при помощи границ областей. Данный метод показывает хорошие результаты на реальных данных [11].

Последним этапом интеллектуального анализа данных на основе алгоритмов Data Mining является решение задачи прогнозирования (Рисунок 1). В предложенном подходе применяются нейронные сети.

Таким образом, актуально применение современных методов искусственного интеллекта для синтеза новых лекарственных препаратов, а так же осуществление сравнительного анализа алгоритмов в зависимости от исходного набора дескрипторов химических соединений.

Постановка задачи исследований

Необходимо провести исследование применения методов Data Mining для решения задачи прогнозирования зависимости «структура/свойство» на примере разработанной базы данных дескрипторов противомикробных веществ сульфаниламидной группы с применением современного программного обеспечения для интеллектуального анализа данных Rapid Miner.

Результаты моделирования и экспериментов в среде Rapid Miner

Для решения поставленной задачи рассмотрим группу противомикробных препаратов сульфаниламидной группы. Сульфаниламиды часто используются для исследований в области QSAR при поиске новых лекарственных соединений. Свойства сульфаниламидов описываются на основе дескрипторного подхода. Информация о сульфаниламидах собрана с помощью одной из самых больших баз данных химических веществ Mol-Instincts, которая содержит более 2.85 млн. химических веществ, 2100 наборов данных по компонентам и в общей сложности более 10 биллионов наборов данных химической информации. Данные ресурса Mol-Instincts хранятся в формате .csv (Comma-Separated Values), который легко конвертируются в другие форматы, например .arff (Attribute-Relation File Format) позволяющий работать с мощными инструментами Data Mining, WEKA (Waikato Environment for Knowledge Analysis), Rapid Miner, RStudio и др.

В таблице 1 представлен фрагмент разработанной базы данных дескрипторов сульфаниламидов. Где D1 – число атомов (number of atoms); D2 – относительное число C атомов (Relative number of C atoms); D3 - относительное число H атомов, (Relative number of H atoms); D4 - относительное число O атомов (Relative number of O atoms); D5 - относительное число N атомов (Relative number of N atoms) и т.д.

Таблица 1- Фрагмент базы данных дескрипторов сульфаниламидов

Дескриптор	Sulfadiazine	Sulfadimidine	Sulfafurazole	Sulfamethizole	...	Sulfaperin
D1	27.0000	33.0000	31.0000	27.0000	...	30.0000
D2	0.370400	0.363600	0.354800	0.333300	...	0.366700
D3	0.370400	0.424200	0.419400	0.370400	...	0.400000
D4	0.074100	0.060600	0.096800	0.074100	...	0.066700
D5	0.148100	0.121200	0.096800	0.148100	...	0.133300
D6	0.037000	0.030300	0.032300	0.074100	...	0.033300
D7	16.0000	24.0000	26.0000	23.0000	...	22.0000
D8	0.571400	0.705900	0.812500	0.793100	...	0.709700
D9	0.00	0.00	0.00	0.00	...	0.00
D10	0.00	0.00	0.00	0.00	...	0.00
D11	12.0000	10.0000	6.0000	6.0000	...	9.0000
D12	0.428600	0.294100	0.187500	0.206900	...	0.290300
...	...	...	...	...	...	...
D2005	250.2751	278.3287	267.3018	270.3240	...	264.3018

В качестве программного обеспечения используется пакет прикладных программ Rapid Miner, который позволяет разрабатывать графические модели и содержит более 400 операторов и алгоритмов Data Mining.

Рассмотрим решение задачи предварительной обработки базы данных дескрипторов сульфаниламидов с помощью алгоритма Random Forest. Фрагмент листинга построения деревьев решений представлен ниже:

```
number of rotatable bonds >
3.500

|   number of multiple bonds
> 13.500: Long_acting
{short_acting=0, medium_act-
ing=0, Long_acting=2}
|   number of multiple bonds
≤ 13.500
|   |   number of multiple
bonds > 8.500
|   |   |   number of multi-
ple bonds > 11.500:
short_acting {short_acting=3,
medium_acting=0, Long_act-
ing=0}
|   |   |   number of multi-
ple bonds ≤ 11.500: Long_act-
ing {short_acting=0, me-
dium_acting=0, Long_acting=2}
|   |   number of multiple
bonds ≤ 8.500: short_acting
{short_acting=3, medium_act-
ing=0, Long_acting=0}
number of rotatable bonds ≤
3.500:
medium_acting {short_act-
ing=1, medium_acting=4,
Long_acting=0}
```

Визуализация ансамбля деревьев решений в среде Rapid Miner показана на Рисунке 2.

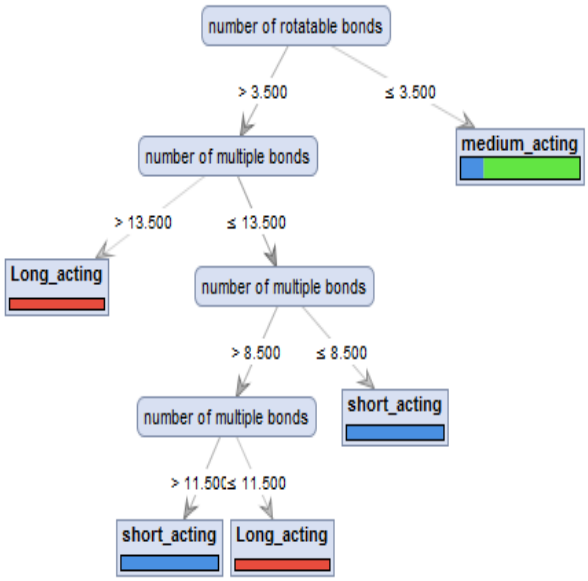

Рисунок 2 – Построение модели Random Forest в пакете прикладных программ Rapid Miner

На Рисунке 3 представлена визуализация базы данных дескрипторов сульфаниламидов по трем веществам в среде Rapid Miner. Графическая модель, построенная на базе программного обеспечения Rapid Miner позволяет реализовать процесс предварительной обработки данных с помощью блока оператора Random Forest и специального инструмента для ранжирования переменных по степени значимости Weight by tree Importance (Рисунок 4).

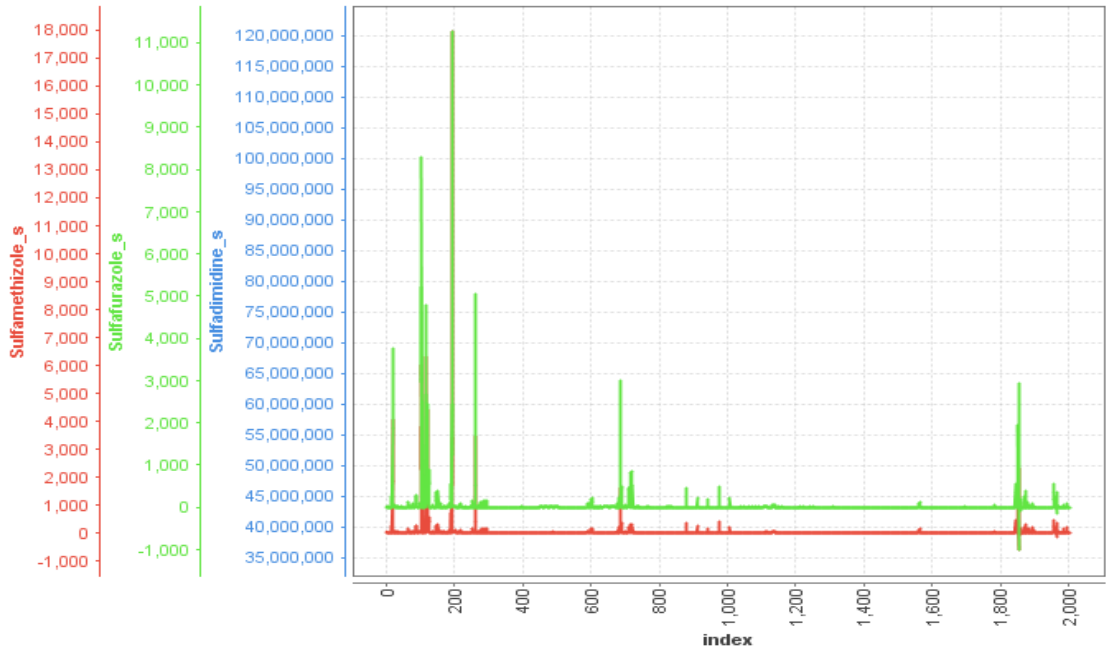

Рисунок 3 - Визуализация дескрипторов сульфаниламидов в среде Rapid Miner

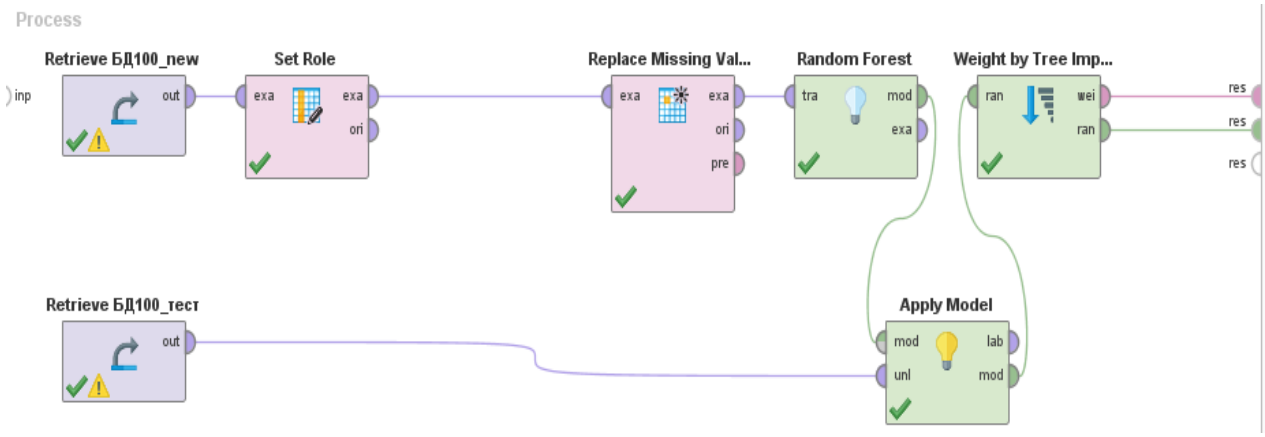

Рисунок 4 – Графическая модель предварительной обработки данных сульфаниламидов в среде Rapid Miner на основе алгоритма Random Forest

На рисунке 5 показаны результаты ранжирования дескрипторов сульфаниламидов по степени значимости. Чем больше вес (weight) дескриптора, тем наиболее ценным он является для дальнейших исследований.

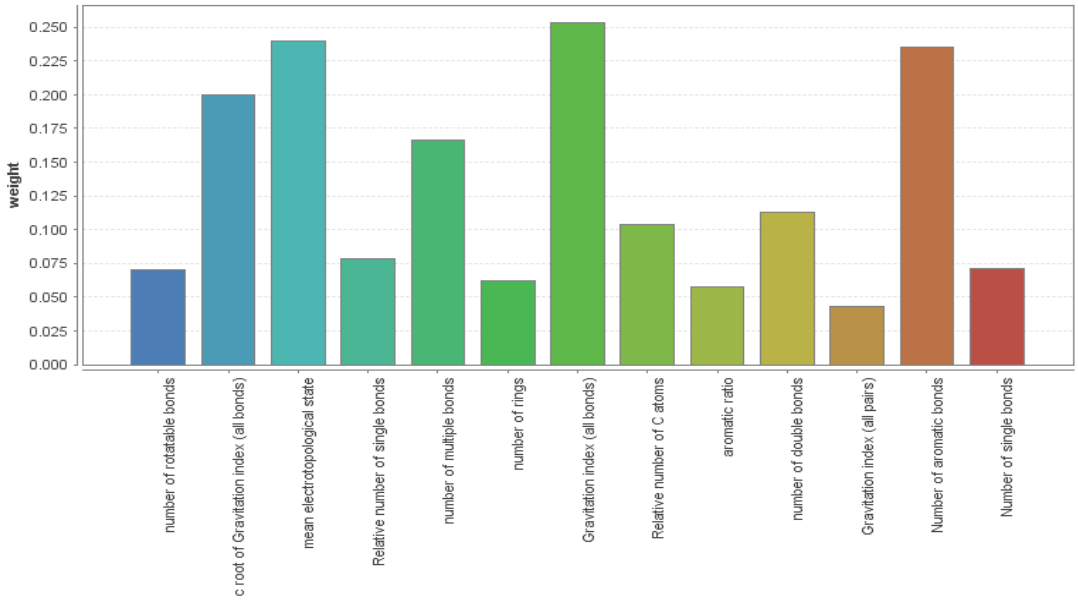

Рисунок 5 – Ранжирование дескрипторов сульфаниламидов по степени значимости в среде Rapid Miner

Далее рассмотрим применение статистического метода для предварительной обработки дескрипторов сульфаниламидов, метода главных компонент (Principle Component Analysis, PCA). Построение графическое модели в среде Rapid Miner осуществляется с помощью блока оператора PCA и блока для построения матрицы корреляции (Рисунок 6).

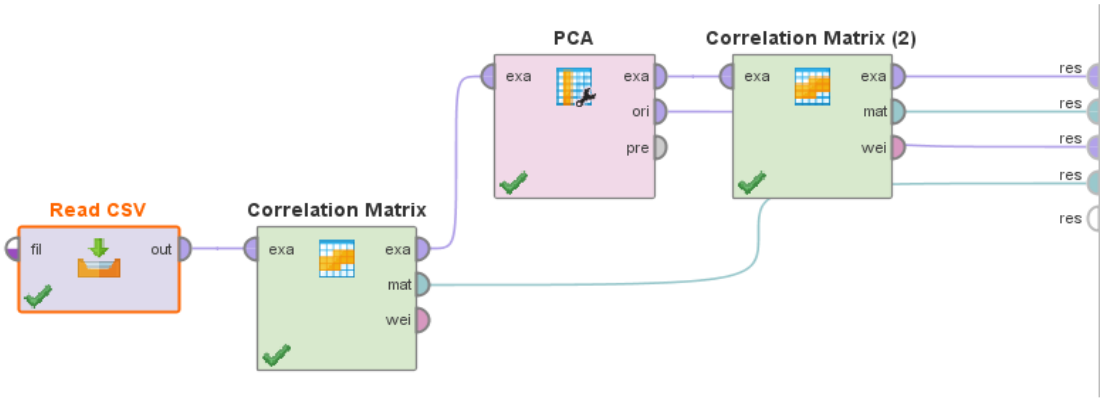

Рисунок 6 – Графическая модель предварительной обработки дескрипторов сульфаниламидов с помощью метода главных компонент

Процедура факторного анализа сводится к следующим этапам: вся совокупность данных поворачивается против часовой стрелки таким образом, что первая ось ассоциируется с максимум дисперсии, а каждая последующая с остаточной. В результате поворота совокупности данных, признаки, которые лежат ближе к началу координат являются малоинформативными и подлежат редукции. На рисунке 7 представлены результаты моделирования в 3D форме.

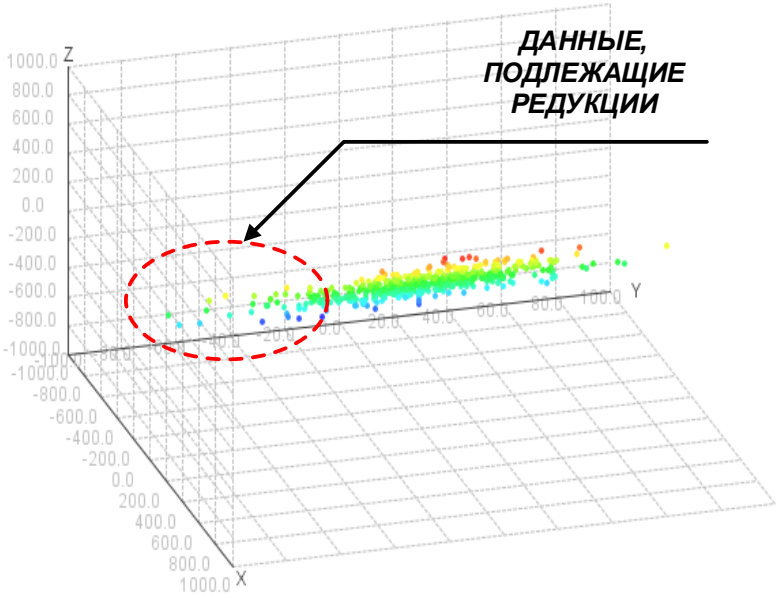

Рисунок 7 – Визуализация в 3D форме результатов моделирования с помощью метода главных компонент дескрипторов сульфаниламидов в среде Rapid Miner

Таким образом, полученные результаты моделирования дескрипторов сульфаниламидов после предварительной обработки данных на основе алгоритма Random Forest и метода главных компонент могут быть использованы в дальнейшем для прогнозирования зависимости «структура/свойство» химических соединений. В результате проведенных исследований выбирается алгоритм, который имеет наименьшую ошибку в зависимости от характера обрабатываемых данных.

Работа выполнена по гранту №ГР 0115РК00549 МОН РК по теме: Компьютерный молекулярный дизайн лекарственных препаратов на основе иммунносетевого моделирования (2015-2017 гг.).

Литература

[1] *Montañez-Godínez N.; Martínez-Olguín A.C.; Deeb O.; Garduño-Juárez R.; Ramírez-Galicia G.* QSAR/QSPR as an Application of Artificial Neural Networks //Artificial Neural Networks. -2014. - №1260. – P. 319-333. **[2]** *Fernandes M.; Caballero J.; Fernandez L.; Sarai A.* Genetic algorithm optimization in drug design QSAR: Bayesian-regularized genetic neural networks (BRGNN) and genetic algorithm-optimized support vectors machines (GA-SVM) // Molecular Diversity. – 2011. - №1. – P. 269-289. **[3]** *Ivanciuc O.* Drug Design with Artificial Intelligence Methods // Encyclopedia of Complexity and Systems Science. – 2009. – P.2113-2139. **[4]** *Samigulina G.A.; Samigulina Z.I.* Immune Network Technology on the Basis of Random Forest Algorithm for Computer-Aided Drug Design //Bioinformatics and Biomedical Engineering. – Granada, Spain, 2017. –Vol. 1. - P. 50-61. **[5]** *Darnag R.; Mostapha Mazouz E.L.; Schmitzer A.; Villemin D.; Jarid A.; Cherqaoui D.* Support vector machines: Development of QSAR models for predicting anti-HIV-1 activity of TIBO derivatives // European Journal of Medicinal Chemistry. 2010. – Vol.45. - №4. – P. 1590 – 1597. **[6]** *Sahigara F.; Ballabio D., Todeschini R., Consonni V.* Defining a novel *k*-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions. – 2013. –Vol.5.- №27. – P. **[7]** *Hemmateenejad B.; Miri R.; Elyasi M.* A segmented principal component analysis—regression approach to QSAR study of peptides // Journal of Theoretical Biology. – 2012. – Vol. 305. – P. 37 – 44. **[8]** *Teixeira A.; Leal J.; Falcao1 A.* Random forests for feature selection in QSPR Models - an application for predicting standard enthalpy of

formation of hydrocarbons // Journal of Cheminformatics. – 2013. – Vol.5. - №9. – P. 1-15.
[9] *Breiman L.* Random Forest //Machine Learning. – 2001. – Vol. 45, №1. – P.5-32.
[10] *Чубукова И.А.* DataMining. – М.: Бином, 2008. – 382с. **[11]** *Bishop C.M.* Pattern Recognition and Machine Learning. – NY: Springer Science + Business Media, 2006. – P.758

Принята впечать 16.07.17

Самигулина Г.А., Самигулина З.И.
Институт информационных и вычислительных технологий,
Алматы, Казахстан

ПРИМЕНЕНИЕ СОВРЕМЕННЫХ МЕТОДОВ DATA MINING ДЛЯ ПРОГНОЗИРОВАНИЯ ЗАВИСИМОСТИ «СТРУКТУРА/СВОЙСТВО» ХИМИЧЕСКИХ СОЕДИНЕНИЙ СУЛЬФАНИЛАМИДОВ

Аннотация. В настоящее время актуальна разработка современных интеллектуальных технологий прогнозирования новых лекарственных препаратов с заданными фармакологическими свойствами на базе методов искусственного интеллекта и статистического анализа данных. Статья посвящена исследованиям в области компьютерного молекулярного дизайна новых лекарственных препаратов сульфаниламидов с применением современного программного обеспечения для интеллектуального анализа данных Rapid Miner. Разработана база данных дескрипторов сульфаниламидов на основе крупнейшего мирового ресурса химической информации Mol-Instincts. Представлена графическая модель предварительной обработки дескрипторов сульфаниламидов в среде Rapid Miner с помощью алгоритма Random Forest и метода главных компонент. Получены результаты численного моделирования, осуществлена визуализация данных в 2D и 3D форме.

Ключевые слова: data mining, сульфаниламидтер дескрипторы

G.A. Samigulina G.A., Z.I. Samigulina Z.I.
Institute of Information and Computing Technologies,
Almaty, 050010, str. Pushkeen 125, Kazakhstan
e-mail: galinasamigulina@mail.ru, zarinasamigulina@mail.ru

APPLYING OF DATA MINING METHODS FOR PROGNOSYS OF "STRUCTURE/PROPERTY" DEPENDENCE OF SULFONAMIDES CHEMICAL COMPOUND

Abstract. Nowadays, is very actual the development of modern intellectual technologies for predicting new drugs with prescribed pharmacological properties based on artificial intelligence methods and statistical data analysis. The article is devoted to researches in the field of computer molecular design of sulfanilamide new drugs with the use of modern intellectual data analysis software RapidMiner. There was developed a database of sulfonamides descriptors based on the world's largest chemical information resource Mol-Instincts. Was presented a graphical model of preliminary processing of sulfonamides descriptors in the RapidMiner environment using the RandomForest algorithm and the principal component method. Were obtained the results of numerical modeling, were made the visualization of data in 3D form.

Keywords: data mining, sulfanilamide descriptors

Ақпараттық және есептеуіш технологиялар институты,
Алматы қ.050010, Пушкин көш. 125,
e-mail: galinasamigulina@mail.ru, zarinasamigulina@mail.ru

СУЛЬФАНИЛАМИДТЕРДІҢ ХИМИЯЛЫҚ ҚОСЫЛЫСТАРЫНЫҢ «ҚҰРЫЛЫМЫ/ҚАСИЕТІ» ТӘУЕЛДІЛІКТЕРІН БОЛЖАУ ҮШІН DATA MINING ЗАМАНУИ ӘДІСІН ҚОЛДАНУ

Аннотация. Қазіргі таңда деректерді статистикалық талдау мен жасанды интеллект әдістері негізінде жаңа дәрілік препараттардың берілген фармакологиялық қасиеттерін болжаудың заманауи интеллектуалды технологияларын жасау өзекті мәселе болып тұр. Мақала RapidMiner деректерін жасанды талдауға арналған заманауи бағдарламалық қамтамасыз етуді қолдана отырыпсульфаниламидті жаңа дәрілік препараттарының компьютерлік молекулалық дизайнын зерттеуге арналған. Әлемдік орасан зор Mol-Instincts химиялық ақпараттар қорының негізіндегі сульфаниламидтердің дескипторларының дерекқоры жасалды. Басты компоненттері мен Random Forest алгоритмінің көмегімен Rapid Miner ортасында сульфаниламидтердің дескрипторларын алдын ала өңдеудің графикалық моделі ұсынылды. Сандық модельдеудің нәтижесі алынып, 3D формасында деректерді визуализациялау жүзеге асырылды.

Түйінсөздер: data mining, сульфаниламидтер дескрипторы