УДК 004.93 + 004.85

ДВУХФАЗНАЯ СХЕМА РЕШЕНИЯ В РАМКАХ ИСПОЛЬЗОВАНИЯ СМЕСЕЙ АЛГОРИТМОВ В ЗАДАЧЕ «СТРУКТУРА – СВОЙСТВО» [1]

Прохоров Е.И. [1], Свитанько И.В.[2], Захаренко А.Л.[3], Суханова М.В.[3]., Беккер А.В.[1], Перевозников А.В.[1], Кумсков М.И.[1]

[1]Московский государственный университет им. М.В. Ломоносова,

[2] Высший химический колледж РАН

[3] Институт химической биологии и фундаментальной медицины СО РАН

E-mail: eugeny.prokhorov@gmail.com, qsar_msu@mail.ru

Статья посвящена прогнозированию свойств химических соединений математическими методами распознавания образов. Исследование проведено на примере активности ингибиторов фермента деления клеток. В качестве методов построения распознающих моделей используется подход на базе смесей алгоритмов. В работе рассмотрена двухфазная схема решения задачи «структура – свойство», также описаны локальный классификатор на базе метода ближайших соседей и метод использующий множества кластеризаций. Проведено сравнение новых алгоритмов построения смесей классификаторов. Рассматриваются методы согласованного прогнозирования активности новых соединений. Также приводится сравнение результатов математического моделирования с методами молекулярного дизайна на основе координации соединений с известными структурами терапевтических мишеней. Проведено экспериментальное изучение биологической активности.

Ключевые слова: смеси алгоритмов, классификация, распознавание образов, кластеризация, задача «структура – свойство», QSAR (Quantitative Structure-Activity Relationship), докинг

Введение

Рассматриваемая в работе задача «структура – свойство» состоит в выявлении зависимостей между структурой химического соединения и его физико-химическими свойствами (QSPR), а также биологической активностью (QSAR). Задача «структура – свойство» является актуальной задачей классификации и для её решения в настоящий момент активно используются математические методы распознавания образов [1]. Ключевой особенностью задачи «структура – свойство» является её ориентированность на предсказание активности новых неизученных соединений. Ввиду ограниченности

[1] Работа выполнена при финансовой поддержке гранта РФФИ № 10-07-00694, 12-03-01036-а и 12-03-92420.

мощности обучаемой выборки на практике, предсказание активности произвольного химического соединения не представляется возможным, поэтому для прикладного использования распознающих моделей является необходимым реализовывать ограничения допустимости для химических соединений [2, 3]. Таким образом, для задачи «структура – свойство» является важным решить задачу отказа от прогноза, известную также как задачу обнаружения нетипичных объектов или «новизны» (novelty detection).

Средствами для решения поставленной задачи обладает математическая теория распознавания образов. Мы сосредоточим внимание в первую очередь на методах, получивших название композиции классификаторов. Наиболее общее и классическое изложение сути алгоритмической композиции дано Ю.И. Журавлевым в алгебраическом подходе [4]. В дальнейшем подход получил активное развитие, как в зарубежных публикациях, так и в отечественных [5]. Более узко проблему можно решать в рамках подхода, называемого смеси алгоритмов.

Смеси алгоритмов позволяют реализовать прогнозирование локально. В русскоязычной литературе эта идея основана на понятии области компетентности, введенным Растригиным в [6]. В общем же случае концепция смеси алгоритмов включает в себя три ключевые компоненты [7]: набор классификаторов (базовых алгоритмов), набор шлюзов, разбивающих исходное пространство объектов на нечеткие области, в которых компетентно мнение каждого из классификаторов, а также вероятностную модель, объединяющую шлюзы и классификаторы. Модель представляет собой взвешенную сумму ответов классификаторов с весами равными значению соответствующих шлюзов.

В такой простой форме оригинальная концепция смеси алгоритмов обладает тремя важными свойствами:

1) она позволяет отдельным классификаторам специализироваться на более узкой части общей проблемы;
2) она использует нечеткое разбиение исходных данных;
3) она допускает оптимизацию разбиения исходного пространства объектов.

В данной работе в качестве принципиально различных реализаций концепции смеси алгоритмов рассмотрены итеративный алгоритм на базе множества различных кластеризаций обучающей выборки [8], локальный классификатор на базе метода k ближайших соседей (K-NN) [9], а также излагается двухфазная схема решения [10], при которой ограничения допустимости (шлюзы) строятся как решение задачи классификации второго уровня, состоящей в разделении верно и неверно классифицированных соединений моделью первого уровня.

Расчеты проведены для ингибиторов поли-(АДФ-рибоза)-полимеразы-1 (ПАРП) – это фермент, который локализуется в клеточном ядре и катализирует поли-АДФ-рибозилирование различных белков [11]. Ингибирование PARP [11], необходимое для подавления механизмов репарации и выживания клеток при химио- и радио- терапии ,

рассматривается как многообещающая стратегия лечения различных видов рака. Моделирование биологической активности в рамках решения задачи «структура – свойство» представляется на сегодняшний день актуальной задачей молекулярной биологии.

Задача «структура – свойство»

Задача «структура – свойство» разбивается на два этапа, этап описания обучающей выборки и этап поиска функциональной зависимости. Более детальное описание задачи и основные определения приведены в [10]. Здесь же кратко опишем проведенное исследование.

В качестве исходных данных была взята выборка химических соединений, для которой рассматривалась активность – ингибирование фермента деления клеток. В выборке представлено 120 соединений с известной активностью – 86 активных и 34 неактивных. Также представлены 196 молекул с неизвестной активностью.

Задача заключалась в том, чтобы построить модели для прогнозирования активности и применить построенные модели для молекул с неизвестной активностью.

В качестве *дескрипторов* для описания соединений выбран метод выделения линейных фрагментов с введением маркировки вершин молекулярных графов [12].

Регулируемые *параметры описания*:

1) длина линейных фрагментов (k= 2, 3, 4);

2) маркеры, участвующие в описании (d –степень вершины молекулярного графа, b – информация о наличии химических связей, r – положение в кольце).

В соответствии с выбором параметров построено 24 матрицы молекула-дескриптор – 8 вариантов включения маркеров и 3 варианта длины линейных фрагментов. Строкам матрицы соответствуют молекулы выборки, столбцам – дескрипторы. Значение матрицы на пересечении *i*-ой строки и *j*-го столбца – количество повторений *j*-го дескриптора в *i*-ой молекуле.

Для оценки качества моделей использовался функционал качества со скользящим контролем [13]. Напомним, что процедура скользящего контроля (leave-one-out cross-validation) заключается в следующем: из обучающей выборки последовательно удаляется каждое соединение, по оставшимся соединениям строится распознающая модель, и с помощью этой модели прогнозируется активность удаленного соединения. В работе везде будет использован функционал качества моделей на скользящем контроле, равный отношению количества верных прогнозов к общему числу спрогнозированных соединений.

В процессе построения всех обсуждаемых моделей использовался метод эволюционного отбора дескрипторов [14].

В ходе эволюционного отбора дескрипторов для каждой матрицы «молекула дескриптор» производится следующая селективная процедура.

Выбранным алгоритмом распознавания образов строится распознающая модель:

- сначала на МД-матрицах, состоящих только из одного столбца-дескриптора исходной матрицы,
- затем отбираются лучшие в смысле функционала качества со скользящим контролем столбцы,
- формируются новые матрицы, путем последовательного присоединения столбцов-дескрипторов исходной матрицы к лучшим матрицам из предыдущей итерации.

Отбор дескрипторов прекращается, когда присоединение нового столбца уже не дает улучшения функционала качества.

Перейдем теперь к описанию конкретных методов построения смесей алгоритмов.

Множественная кластеризация

В данном варианте построения модели к построенным МД-матрицам применялись различные методы кластеризации (метод минимального покрывающего дерева, иерархический кластерный анализ, метод к-средних) [15] с различными параметрами (количество получаемых кластеров, метрики для вычисления расстояния между объектами и кластерами). С помощью варьирования показателей различных методов кластерации получено 53 варианта кластеризации для каждой из МД-матриц. Примерами различных методов кластеризации могут служить:

1. Кластеризация методом минимального покрывающего дерева с параметром k = 3, 4, 5, 6, равным количеству кластеров, на которое разбивается множество
2. Иерархический кластерный анализ. Реализация данного метода в системе Matlab имеет ряд вариьируемых параметров, таких как метрика для вычисления расстояния между объектами (Евклидова метрика, метрика Хэмминга, метрика Минковского и т.д.) и метод измерения расстояния между кластерами (невзвешенное среднее расстояние, наибольшее/наименьшее расстояние, расстояние от центров масс и т.д.)

Количество вариантов кластеризации неограниченно, но в данном исследовании авторами было выбрано конечное число вариантов разбиения. Следует отметить, что выбор методов кластеризации и их количества зависит исключительно от исследователя и решаемой им задачи.

Таким образом, для каждой из 24 МД-матриц есть 53 варианта разбиения на кластеры и известны результаты построения функции прогноза на этих кластерах. Результаты всех вариантов кластеризации можно представить как множество слоев, где каждый слой соответствует одному из методов кластеризации и представляет собой обучающую выборку с проведенными между объектами границами. Каждый объект x обучающей выборки при таком представлении в каждом слое I принадлежит некоторому кластеру K^i и активность этого объекта может быть прогнозирована с помощью построенной линейной функции $y = f_{K^i}(x)$.

Для построения линейной функции прогноза в работе применялся метод самоорганизации моделей, предложенный Ивахненко А.Г. с соавторами [14], (Метод Группового Учета Аргументов, МГУА), который модифицирован нами для проведения анализа структурных спектров. Использование МГУА позволяет не только отбирать значимые переменные в ходе построения функциональной зависимости, но также дает возможность выполнять функциональные преобразования дескрипторов в ходе расчетов. В качестве результата применения МГУА получаем линейную функцию прогноза $y = f_{K^i}(x)$ и соответствующую ей прогностическую способность $R(f_{K^i})$.

Для каждого фиксированного варианта описания выполним процедуру перестроения кластеров. Алгоритм состоит в следующем: выберем некоторое разбиение $L = \{L_1, L_2, \ldots, L_k\}$ на кластеры, примем его за исходное. Выберем следующий из 53 вариантов разбиения на кластеры. Пусть это будет разбиение $K = \{K_1, K_2, \ldots, K_m\}$. Рассмотрим все пересечения $K \bigcap L$ кластеров исходного и текущего вариантов разбиения. Каждому из пересечений двух кластеров присвоим ту из функций прогноза, которая обладает большей прогностической способностью R.

$$\begin{cases} C = L_i \bigcap K_j \\ R(f_{L_i}) > R(f_{K_j}) \end{cases} \Rightarrow f_c = f_{L_i}$$

Получаем новое разбиение на кластеры. Повторяем алгоритм со следующим вариантом разбиения на кластеры. Процедура останавливается, когда рассмотрены все варианты разбиения на кластеры. В результате мы обрабатываем все результаты кластеризации и построения прогнозирующих функций с целью нахождения функций с наилучшим качеством прогноза. Кроме того, формируются быстрые правила отказа от прогнозирования. В таблице 1 содержатся результаты применения указанного подхода к различным описаниям обучающей выборки.

Маркеры	***	**r	*b*	d**	*br	d*r	db*	dbr
Длина фрагмента	k=2							
Качество прогноза	0.8583	0.9000	0.8583	0.8750	0.9000	0.9417	0.8833	0.9000
Длина фрагмента	k=3							
Качество прогноза	0.9000	0.8833	0.9000	0.8833	0.8833	0.8833	0.9250	0.8833
Длина фрагмента	k=4							
Качество прогноза	0.8500	0.8333	0.8250	0.8250	0.8250	0.8250	0.8000	0.8250

Таблица 1. Качество прогноза моделей на базе множественной кластеризации.

Построение классификатора на базе k-NN

Алгоритм k ближайших соседей (KNN – k nearest neighbors) – один из популярных метрических алгоритмов классификации. Алгоритм основан на так называемой гипотезе

компактности, которая на практике означает, что схожим объектам соответствуют схожие классы или более строго, что классы образуют компактно локализованные подмножества. Для формализации понятия «сходства» в исходном признаковом пространстве вводится функция расстояния.

Классификация объекта *x* основана на множестве объектов обучающей выборки, ближайших к объекту *x* в смысле введенной ранее метрики.

Пусть C_1 множество объектов обучающей выборки, отнесенных к первому классу активности, C_2 множество объектов обучающей выборки, отнесенных ко второму классу активности. Обозначим через C_x множество k ближайших к x объектов обучающей выборки.

Если $|C_x \bigcap C_1| > |C_x \bigcap C_2|$, то объект *x* относим к первому классу активности, в случае $|C_x \bigcap C_1| < |C_x \bigcap C_2|$ объект *x* относим ко второму классу активности. Когда мощности указанных множеств равны (вблизи объекта *x* находится равное число объектов разных классов) ответ классификатора считаем неопределенным.

В настоящей работе использована модификация алгоритма k ближайших соседей с ограничением по радиусу (при классификации не учитываются объекты, находящиеся слишком далеко от объекта классификации). Значение радиуса выбирается на основе результатов кластеризации.

Подробно используемый классификатор описан в работе [9]. Ниже результаты применения локального классификатора на базе KNN к различным описаниям обучающей выборки.

Маркеры	***	**r	*b*	d**	*br	d*r	db*	dbr
Длина фрагмента	k=2							
Качество прогноза	0.8916	0.9500	0.9333	0.9333	0.9833	0.9333	0.9666	0.9833
Длина фрагмента	k=3							
Качество прогноза	0.9333	0.9833	0.9583	0.9666	0.9833	0.9666	0.9750	0.9833
Длина фрагмента	k=4							
Качество прогноза	0.9250	0.9666	0.9333	0.9500	0.9750	0.9666	0.9533	0.9833

Таблица 2. Качество прогноза моделей на базе локального KNN-классификатора.

Двухфазная схема решения задачи «структура – свойство»

В данном разделе рассмотрим пример построения и использования универсальных правил отказа от прогноза для решения задачи «структура – свойство». Предложенный метод описан в [10].

Пусть обучающая выборка *LS* состоит из *N* химических соединений x_i, $i = 1,\dots N$. каждому из которых поставлено в соответствие одно из значений: «1» или «–

1» («1» соответствует активным соединениям, «–1» – неактивным). Вектор, последовательно содержащий активности всех соединений обучающей выборки обозначим $y = (y_1, y_2, \dots y_N)$, $y_i \in \{-1,1\}$.

Пусть также построена распознающая модель RM_1, решающая исходную задачу классификации, то есть $\forall x_i \in LS, \quad RM_1(x_i) \in \{-1,1\}$. RM_1 назовем моделью первого уровня.

Обозначим через R_1 – множество тех соединений обучающей выборки x_i, для которых полученные в ходе процедуры скользящего контроля значения активности совпадают с действительными $RM_1(x_i) = y_i$, то есть множество верно классифицированных моделью первого уровня соединений. Через W_1 обозначим множество ошибочно классифицированных моделью первого уровня соединений – $W_1 = \{x_i \in LS \mid RM_1(x_i) \neq y_i\}$. Таким образом, функционал качества со скользящим контролем для модели первого уровня равен $\varphi_1 = | R_1 | / N$.

Определим задачу классификации второго уровня. Всем соединениям обучающей выборки, спрогнозированным верно моделью первого уровня (таких $| R_1 |$), поставим в соответствие значение «1», а соединениям, спрогнозированным неверно (таких $| W_1 |$) поставим в соответствие значение «-1». Сформируем таким образов вектор $y2 = (y2_1, y2_2, \dots y2_N)$, $y2_i \in \{-1,1\}$.

$$y2_i = \begin{cases} 1, & \textit{если} \quad RM_1(x_i) = y_i, \\ -1, & \textit{если} \quad RM_1(x_i) \neq y_i \end{cases} \quad i = 1, \dots N.$$

Возникшую задачу классификации назовем задачей классификации второго уровня.

Пусть построена распознающая модель RM_2, решающая задачу классификации второго уровня, то есть $\forall x_i \in LS, \quad RM_2(x_i) \in \{-1,1\}$. RM_2 назовем моделью второго уровня. Пусть в ходе процедуры скользящего контроля моделью второго уровня получено $| R_2 |$ верных прогнозов, где $R_2 = \{x_i \in LS \mid RM_2(x_i) = y2_i\}$. Тогда функционал качества модели второго уровня $\varphi_2 = | R_2 | / N$. Наконец определим результирующую распознающую модель RM_0. Результирующая модель решает исходную задачу классификации, но в отличие от модели первого уровня, результирующая модель обладает опцией отказа от прогноза. То есть $\forall x_i \in LS, \quad RM_0(x_i) \in \{-1,0,1\}$ и значение $RM_0(x_i) = 0$ интерпретируется, как отказ от прогноза активности соединения x_i. Для $x_i \in LS$:

$$RM_0(x_i) = \begin{cases} 1, & \textit{если} \quad RM_2(x_i) = 1 \quad \text{и} \quad RM_1(x_i) = 1; \\ -1, & \textit{если} \quad RM_2(x_i) = 1 \quad \text{и} \quad RM_1(x_i) = -1; \\ 0, & \textit{если} \quad RM_2(x_i) = -1. \end{cases}$$

Таким образом, результирующая модель осуществляет отказ от прогноза тогда, когда модель второго уровня предсказывает, что модель первого уровня ошибается, и осуществляет прогноз активности моделью первого уровня в противном случае. Как и ранее, обозначим $R_0 = \{x_i \in LS \mid RM_0(x_i) = y_i\}$ – множество верно классифицированных результирующей моделью соединений. Пусть также через R*eject* обозначено количество отказов от прогноза. Тогда функционал качества результирующей модели $\varphi_0 = \mid R_0 \mid /(N - \text{R}eject)$.

Тогда для оценки качества прогноза при использовании двухфазной схемы решения верна следующая теорема.

Теорема.

$$\varphi_0 = \frac{(\varphi_1 + \varphi_2)N - \text{Reject}}{2(N - \text{Reject})}.$$

Следствием которой является в частности следующее утверждение.

Если $\varphi_2 \geq \varphi_1 > 1/2$, тогда при Reject>0 имеем $\varphi_0 > \varphi_1$.

Также верно, что если $\varphi_2 > \varphi_1 > 1/2$, то $\varphi_0 > \varphi_1$.

Таким образом, доказано, что если модель первого уровня прогнозировала соединения из обучающей выборки, хотя бы чуть лучше, чем случайным образом, и качество модели второго уровня не хуже качества модели первого уровня. Тогда при условии, что количество отказов от прогноза больше нуля, имеем, что результирующая модель демонстрирует более высокое качество классификации на исходной задаче, чем модель первого уровня.

Второе следствие гарантирует улучшение качества прогноза в случае, когда качество модели второго уровня превосходит качество модели первого уровня, не зависимо от числа отказов.

Ниже в таблицах приведены результаты использования данного подхода на регрессионных моделях и моделях на базе метода опорных векторов (SVM) [16]. При использовании SVM применялся Multilayer Perceptron kernel [17]. В таблицах ниже длина линейных фрагментов на этапе описания была фиксирована и равнялась 2. Также для задач классификации первого и второго уровня эволюционный отбор дескрипторов применялся независимо. В таблице ниже столбцы D1 и D2 содержат количество дескрипторов, отобранных для решения задач классификации первого и второго уровня соответственно.

Маркеры	φ_1	φ_2	Отказы	φ_0	D1	D2
***	0,941667	1	7	1	2	1
**r	0,891667	0,883333	1	0,890756	5	1
b	0,908333	0,933333	13	0,971963	3	3
*br	0,941667	0,925	9	0,981982	4	1
d**	0,883333	0,891667	9	0,918919	1	1
d*r	0,875	0,875	8	0,910714	3	1
db*	0,908333	0,925	8	0,946429	2	1
dbr	0,966667	0,925	9	0,981982	4	1

Таблица 3. Качество прогноза двухфазной схемы решения с использованием метода опорных векторов.

Маркеры	φ_1	φ_2	Отказы	φ_0	D1	D2
***	0,841667	0,841667	0	0,841667	3	1
**r	0,808333	0,816667	1	0,815126	4	2
b	0,916667	0,916667	0	0,916667	8	1
*br	0,95	0,958333	1	0,957983	7	2
d**	0,9	0,9	0	0,9	7	1
d*r	0,758333	0,908333	24	0,916667	2	5
db*	0,933333	0,933333	0	0,933333	6	1
dbr	0,95	0,95	0	0,95	5	1

Таблица 4. Качество прогноза двухфазной схемы решения с использованием регрессии.

Прогнозирование активности новых соединений

Построенные по описанным выше алгоритмам модели прогнозирования были применены к выборке соединений с неизвестными значениями активности. Результаты были объединены в сводную таблицу и прогноз активности соединения оценивался по сумме значений активности, полученных на каждой модели. Напомним, что допустимые значения активности в нашем случае «+1» и «–1», соответственно, активное и неактивное соединение.

В результате применения построенных моделей к имеющейся выборке из 196 соединений с неизвестным значением активности можно выделить несколько соединений, предположительно, обладающих активностью.

Наиболее перспективные с точки зрения построенных моделей соединения прошли экспериментальные испытания, результатом которых являлась оценка ингибирующей концентрации. Одно из соединений показало активность. Уточнение результатов дало ингибирующую концентрацию на данном соединении в 200 микроМ (0,2 мМ), то есть соединение является достаточно эффективным ингибитором ПАРП.

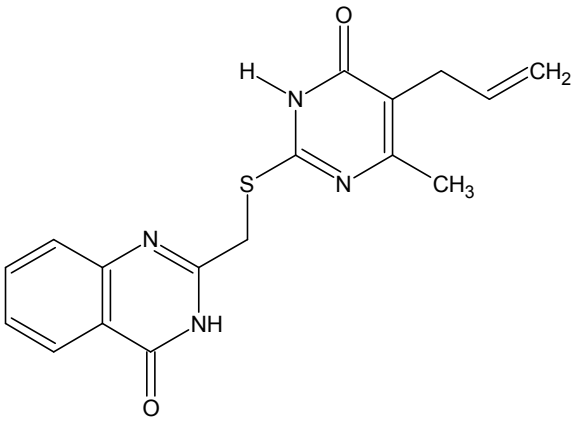

Рис. 1. Структура соединения, показавшего активность на экспериментальных испытаниях.

Для оценки эффективности разработанного метода было проведено его сравнение с наиболее эффективными на сегодняшний день методами моделирования биологической активности химических соединений. Для этого выборка из 196 соединений была также проанализирована методами молекулярного докинга, реализованными в пакете программ Lead Finder [18]. Было показано, что во многих случаях предсказания QSAR-моделей согласуется с предсказаниями молекулярного докинга[19], однако для нескольких соединений, наиболее перспективных с точки зрения построенных моделей методами молекулярного докинга, методом QSAR было предсказано отсутствие биологической активности. Имела место и обратная ситуация.

Таким образом, прогнозирование активности химических соединений на базе QSAR-моделей позволяет устранить существующие ошибки молекулярного моделирования и исключить ложноотрицательные предсказания. Из чего следует однозначные выводы:

1) метод моделирования, описанный в настоящей работе, показывает свою эффективность;

2) два основных метода моделирования биологической активности – QSAR и докинг – дополняют друг друга.

Заключение

В работе изложены различные алгоритмы решения задачи «структура – свойство» в рамках концепции использования смесей алгоритмов. В результате применения различных алгоритмов построено несколько моделей для описания и прогнозирования биологической активности химических соединений.

Все описанные подходы реализованы и применены к конкретной обучающей выборке соединений. На основе полученных результатов построен прогноз значения активности 196 соединений с неизвестной активностью.

Данная работа представляет собой обзор прикладных методов распознавания образов и классификации в области моделирования биологической активности. Полученные результаты говорят о перспективности использования построенных

моделей. Подход, предложенный в данной работе доказал свою эффективность в ходе проведения экспериментальных испытаний.

Литература

1. Прохоров Е.И., Кумсков М.И., Беккер А.В., Перевозников А.В., Пугачева Р.Б., Апрышко Г.Н. Согласованное прогнозирование противоопухолевой активности по семейству моделей «структура-свойство» // Прогнозирование свойств химических соединений. Унифицированный Репозиторий моделей «структура – свойство»: – Сборник научных работ. – М.: МАКС Пресс, 2012. – С. 25-56.

2. E. I. Prokhorov, L. A. Ponomareva, E. A. Permyakov and M. I. Kumskov Fuzzy classification and fast rules for refusal in the QSAR problem // Pattern Recognition and Image Analysis, 2011, Volume 21, Number 3, Pages 542-544

3. Прохоров Е.И. «Нечеткое» прогнозирование свойств химических соединений: Использование нечеткой функции классификации на кластерах обучающего множества в задаче «структура – свойство», Saarbrucken, Germany: LAP Lambert Academic Publishing, 2012, – 80 с.

4. Журавлёв Ю. И. Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. 1978. Т. 33. С. 5–68.

5. Воронцов К.В., Каневский Д.Ю., «Коэволюционный метод обучения алгоритмических композиций» // Таврический вестник информатики и математики, 2005, № 2, 51–66.

6. Растригин Л. А., Эренштейн Р. Х. Коллективные правила распознавания. М.: Энергия, 1981. – 244 с.

7. Adaptive mixtures of local experts / R. A. Jacobs, M. I. Jordan, S. J. Nowlan, G. E. Hinton // Neural Computation. 1991. no. 3. Pp. 79–87.

8. Беккер А.В. Применение кернел-методов в задаче «структура – свойство». Прогнозирование свойств и биологической активности химических соединений, Saarbrucken, Germany: LAP Lambert Academic Publishing, 2012, – 96 с.

9. A. V. Perevoznikov, A. M. Shestov, E. A. Permyakov, M. I. Kumskov A Way to Increase the Prediction Quality for the Large Set of Molecular Graphs by Using the k_NN Classifier // Pattern Recognition and Image Analysis. – 2011. – Vol. 21; No. 3, pp. 545–548.

10. Прохоров Е.И. Нейронные сети для построения ограничений допустимости в задаче «структура – свойство» // Нейрокомпьютеры: разработка, применение. 2012. № 10. 46–56.

11. D.D' Amours, S. Desnoyers, I. D'Silva and G. G. Poirier Poly(ADP-ribosyl)ation reactions in the regulation of nuclear functions. // Biochem. J. 1999. 342 (Pt 2). 249–268.

12. Кумсков М.И., Смоленский Е.А., Пономарева Л.А., Митюшев Д.Ф., Зефиров Н.С. Системы структурных дескрипторов для решения задач «структура – свойство» // Доклады АН, 1994, т.336, N1, с.64–66.

13. Stone M. Cross–Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society, B, 36, pp. 111–147, 1974.

14. Кумсков М.И., Митюшев Д.Ф. Применение метода группового учета аргументов для построения коллективных оценок свойств органических соединений на основе индуктивного перебора их «структурных спектров». // Проблемы управления и информатики, 1996, №4, с.127–149.

15. P. Berkhin Survey of Clustering Data Mining Techniques, Accrue Software, 2002.

16. Vapnik, V.N. The nature of statistical learning theory. V.N. Vapnik. New York; London: Springer. 1998.

17. Thomas R., Karsten B. Multilayer Perceptron kernel. // Proceedings of the 24th SIBGRAPI Conference on Graphics, Patterns and Images. Maceio, Alagoas, Brazil. 2011. 337–343.

18. Stroganov O.V., Novikov F.N., Stroylov V.S., Kulkov V., Chilov G.G. Lead finder: an approach to improve accuracy of protein-ligand docking, binding energy estimation, and virtual screening // J Chem Inf Model. 2008 Dec; 48(12):2371–85.

19. Leonid V. Romashov, Alexey A. Zeifman, Alexandra L. Zakharenko, Fedor N. Novikov, Viktor S. Stroilov, Oleg V. Stroganov, Germes G. Chilov, Svetlana N. Khodyreva, Olga I. Lavrik, Ilya Yu. Titov and Igor V. Svitan'ko. Rational design and synthesis of new PARP1 inhibitors. Mendeleev Communications, 22(1), 15-17 (2012).

Прохоров Евгений Игоревич, к.ф.-м.н. (теоретические основы информатики), окончил специалитет (2010 г.) и аспирантуру (2013 г.) механико-математического факультета МГУ имени М.В. Ломоносова (кафедры вычислительной математики). Автор 15 публикаций в области прогнозирования свойств химических соединений. Область научных интересов – машинное обучение, прогнозирование свойств химических соединений, классификация.

Свитанько Игорь Валентинович, Московский государственный университет им. М.В. Ломоносова, 1978 г. Институт органической химии им. Н.Д. Зелинского РАН, 1978-настоящее время. Кандидат наук с 1983 г., МГУ имени М.В. Ломоносова, Химический факультет, Кафедра фундаментальных проблем химии, доцент, с 2009 г. Область интересов: направленный синтез, биологически активные соединения, докинг. Опубликовано 46 научных работ, 5 глав в монографиях, 7 учебников.

Захаренко Александра Леонидовна

Родилась в 1973 году в Новосибирске; закончила ФЕН НГУ по специальности «химия» в 1990 году, год защиты кандидатской диссертации – 2001. Научный сотрудник Института химической биологии и фундаментальной медицины СО РАН. Круг научных интересов – репарация ДНК, ингибиторы ферментов репарации. 14 публикаций, 2 патента.

Суханова Мария Владиславовна.

Родилась в 1973 году. В 1990 г. закончила ФЕН НГУ по специальности «биология», в 2008 году защитила кандидатскую диссертацию. Старший научный сотрудник Института химической биологии и фундаментальной медицины СО РАН. Круг научных интересов – репарация ДНК. 21 публикация, 1 патент.

Беккер Александра Владимировна, окончила магистратуру механико-математического факультета
МГУ имени М.В. Ломоносова в 2012 году. Область научных интересов: распознавание образов, прогнозирование свойств химических соединений, data mining.

Перевозников Александр Владимирович, окончил аспирантуру механико-математического факультета
МГУ имени М.В. Ломоносова каф. вычислительной математики в 2013 году. Окончил механико-математический факультет МГУ имени М.В.Ломоносова в 2010 году. Автор 9 статей. Область научных интересов - распознавание образов в химии, прогнозирование свойств химических соединений.

Кумсков Михаил Иванович, д.ф.-м.н., проф. механико-математического факультета МГУ имени М.В. Ломоносова и факультета компьютерных наук НИУ ВШЭ. Окончил в 1978 году ф-т Вычислительной математики и кибернетики МГУ имени М.В. Ломоносова. Автор более 80 публикаций. Область научных интересов – распознавание образов в химии, прогнозирование свойств химических соединений.