ЖУРНАЛ ОБЩЕЙ ХИМИИ, 2020, том 90, № 10, с. 1583–1600

УДК 544.35;544.341.2;519.25

РАЗВЕДОЧНЫЙ, РЕГРЕССИОННЫЙ И НЕЙРОСЕТЕВОЙ АНАЛИЗ УСТОЙЧИВОСТИ КОРОНАТОВ КАТИОНОВ В НЕКОТОРЫХ ЧИСТЫХ РАСТВОРИТЕЛЯХ

© 2020 г. Н. В. Бондарев*

Харьковский национальный университет имени В. Н. Каразина, пл. Свободы 4, Харьков, 61022 Украина
**e-mail: bondarev_n@rambler.ru*

Поступило в Редакцию 13 мая 2020 г.
После доработки 29 июля 2020 г.
Принято к печати 9 августа 2020 г.

Проведен разведочный, регрессионный и нейросетевой анализ констант устойчивости комплексов краун-эфиров [12C4, 16C5, $(CH_3)_2$16C5, DB21C7, DB24C8, DCH24C8, DB30C10] с катионами щелочных (Li^+, Na^+, K^+, Cs^+, Rb^+), щелочноземельных (Mg^{2+}, Ca^{2+}, Sr^{2+}, Ba^{2+}), тяжелых (Ag^+, Tl^+, Co^{2+}, Cu^{2+}, Pb^{2+}) металлов и NH_4^+ состава 1:1 в воде и органических растворителях (метаноле, ацетонитриле, ацетоне, *N*,*N*-диметилформамиде, нитробензоле, нитрометане, 1,2-дихлорэтане, пропиленкарбонате) при 298.15 K, полученных кондуктометрическим методом. Разработаны факторная, кластерные, дискриминантная, каноническая, дерево решений, регрессионные и нейросетевые модели кластеризации, аппроксимации и прогнозирования термодинамических констант комплексообразования краун-эфиров с катионами в зависимости от свойств лиганда, взаимодействующего с ним катиона и используемого растворителя. Обученный многослойный персептрон-кластеризатор MLP 7-5-5 на сто процентов подтвердил кластеризацию, проведенную разведочным методом *k*-средних. На независимых данных по константам устойчивости коронатов демонстрируются прогностические возможности обученного персептрона-аппроксиматора MLP 7-7-1.

Ключевые слова: краун-эфиры, константа комплексообразования, разведочный анализ, множественная линейная регрессия, нейронные сети, моделирование, прогнозирование

DOI: 10.31857/S0044460X20100145

На семинаре, посвященном столетию Дж. Тьюки [1], было отмечено, что в свое время один из создателей практического анализа данных призвал к реформированию академической статистики и указал на существование пока еще непризнанной науки, предметом интереса которой является изучение данных или «анализ данных» [2]. Исходные концепции и принципы Дж. Тьюки не утратили своего значения и формируют часть фундамента современной науки о данных [3].

Работы, выполненные исследователями в разных областях науки и техники за последние десятилетия, являются ярким тому подтверждением. Прежде чем анализировать наборы данных, Д. Чамберс [4] рекомендует уделять больше внимания подготовке и представлению данных, а Л. Брейман [5] призывает делать основной упор на прогностический потенциал математических моделей, а не на умозаключения.Всестороннее, почти энциклопедическое описание статистических методов и аналитических подходов, используемых в науке, промышленности, бизнесе и интеллектуальном анализе данных, представленных с точки зрения практического специалиста («потребителя») этих методов, содержится в работе [6].

Для многих областей науки и техники, нуждающихся в анализе больших массивов данных, особенно в здравоохранении, экологии, химии, биологии, медицине и науке о Земле, прогнозирующее моделирование и машинное обучение предостав-

ляют беспрецедентные возможности для открытия новых знаний и развития теории [7].

Разработаны технологии неконтролируемого (классификация без обучения) и контролируемого (классификация с обучением) распознавания образов, включая анализ главных компонентов (PCA), алгоритм ближайших соседей (NN), дискриминантный анализ частичных наименьших квадратов (PLS-DA) и искусственную нейронную сеть (ANN) [8].

Разведочный анализ является важным шагом после сбора данных и предварительной их обработки во многих типах исследований, но особенно полезен при анализе электронных медицинских записей [9].

Открытый обмен данными, совместное использование наборов данных, метаданных, моделей, программного обеспечения и других ресурсов для анализа повышает точность, достоверность и воспроизводимость результатов, ведет к конструктивистским подходам в науке и способствует экономическому сотрудничеству и развитию [10].

Представлен [11] совместный опыт геологов и IT-специалистов по использованию визуальных разведочных методов анализа данных для изучения закономерностей связи между свойствами химических элементов и минеральных веществ, создания гипотез в области наук о Земле.

Предложены средства контроля качества фармацевтической продукции, основанные на методах спектроскопии, в основном ближнего инфракрасного диапазона, в сочетании с хемометрическими алгоритмами [12]. Обзор основных методов классификации показателей качества продуктов питания, представленных в хемометрической литературе, приведен в работе [13]. Контролируемому моделированию многомерных данных в аналитической химии – построению моделей аппроксимации и дискриминации, их количественной валидации для успешного применения на практике – посвящена работа [14].

В обзоре [15] обсуждаются возможности и универсальность хемометрических методов в свете проблем с большими массивами (био)химических данных, которые встречаются в хроматографии и спектроскопических исследованиях, с акцентом на их применении к «-омика» наукам (геномика, транскриптомика, протеомика или метаболомика).

Разведочный анализ данных применялся [16] для изучения поведения радиоактивных аэрозолей, присутствующих в приземной атмосфере Гранады, с использованием радиоактивного ^{7}Be. Авторы работы [17] применили методы разведочного анализа данных для оценки сходства и кластеризации хиральных полисахаридых систем, используемых для разделения фармацевтических препаратов в жидкостной хроматографии.

Искусственные нейронные сети, как один из самых популярных алгоритмов машинного обучения, широко применяются в различных областях. Объединение знаний в области химии с машинным обучением (анализ данных, нейросетевые прогнозы, мониторинг химических систем) способствует [18]: познанию природы химических веществ, рациональному планированию экспериментов, созданию новых материалов и технологий, зарождению новых концепций химии.

Развиты нейросетевые алгоритмы [19], использующие новый метод дактилоскопии органических реакций. Построена умная (smart) система, которая, учитывая набор реагентов и реактивов, предсказывает вероятные продукты химического превращения.

Авторы [20] разрабатывают нейронную сеть, обучаемую методом обратного распространения на основе информации о химических реакциях. Нейросеть реализована в моделируемой химической системе, где нейроны отделены друг от друга полупроницаемой клеточной мембраной.

Представлен контролируемый подход к обучению граф-сверточной нейронной сети для предсказания продуктов органических реакций по свойствам реагентов, реактивов и растворителей [21]. Показано, что искусственные нейронные сети являются мощной альтернативой традиционным методам оценки степени восстановления деградированной почвы в зависимости от ее химических и физических свойств [22].

Представлены два метода классификации лекарств: классификация сверточными нейронными сетями по химической структуре и классификация случайными лесами по молекулярным отпечаткам пальцев, которые превзошли по эффективности предыдущие прогнозные модели [23].

В обзоре [24] обобщены результаты применения искусственных нейронных сетей для иссле-

дования и прогнозирования катализа, понимания природы каталитических процессов и структур новых катализаторов. В работе [25] представлен эффективный подход глубокого обучения на основе тензорных нейронных сетей, позволяющий понять пространственные и химические особенности квантово-химических молекулярных систем. Разработана стратегия прогнозирования вязкости ненасыщенных сложных полиэфиров (полиэфирных смол) с помощью нейронных сетей. Благодаря нейронным сетям, разработка новых экологически чистых реактивных разбавителей может быть ускорена [26].

В статье [27] сравниваются и обсуждаются результаты прогнозирования температуры стеклования полимеров алгоритмами искусственной нейронной сети и линейной множественной регрессии для создания вычислительных систем для разработки составов полимерных материалов, в том числе полимерных оптических волокон, с желаемыми потребительскими характеристиками. Прогнозированию стабильности кристаллов с помощью глубинных нейронных сетей, для обучения которых используются только два дескриптора – электроотрицательность Полинга и ионные радиусы, посвящена работа [28].

Обсуждаются проблемы, связанные с машинным обучение в области материаловедения, предлагаются возможные решения и перспективные направления будущих исследований по созданию новых материалов [29]. Разработан [30] мягкий датчик для определения содержания этанола в продуктах периодической дистилляции арбузного вина, исходя из температуры кипения. Построенная модель состоит из многослойной персептронной искусственной нейронной сети с одним скрытый слоем. В работе [31] представлен обзор и анализ самых последних исследований, которые развивают и применяют машинное обучение, в частности нейросетевой анализ, к твердотельным системам для открытия стабильных материалов и прогнозирования их кристаллической структуры.

На основе искусственных нейронных сетей в работе [32] предлагается новый подход для универсального описания термодинамических функций чистых веществ в широком температурном диапазоне (от 0 до 6000 K).

Демонстрируется [33–37] применение разведочных методов анализа данных и нейросетевых алгоритмов для группирования пациентов, принимавших бензодиазепины, по медицинским показателям [33]; для кластеризации, аппроксимирования и прогнозирования силы слабых органических кислот [34, 35] и устойчивости коронатов катионов натрия и калия [36, 37] в водно-органических растворителях.

Следует отметить, что в физикохимии растворов, как и в химии в целом, накоплено огромное количество экспериментальных данных, проведение глубокого анализа которых уже невозможно без применения средств современной информатики – «науки о принципиально новой человеко-машинной технологии расширенного воспроизводства качественно нового знания» [38].

Узкое, но очень распространенное понимание хемоинформатики подразумевает применение методов информатики в биоорганической химии для создания лекарств [39]. В дальнейшем эта дефиниция была расширена. В частности, согласно определению, данному Г. Пэризом (2000), хемоинформатика – это научная дисциплина, охватывающая дизайн, создание, организацию, управление, поиск, анализ, распространение, визуализацию и использование химической информации [40], в предмет исследования которой включены приемы хранения, извлечения и обработки химической информации.

Развитию хемоинформатики в значительной мере способствует наличие обоснованной методологии анализа данных и реализующего ее программного обеспечения (STATISTICA, SPSS, R, SAR/QSAR/QSPR), которые позволяют химику на основе обработки экспериментальных данных осуществлять прогнозирование самых разнообразных свойств химических соединений и процессов [34–37, 41–44]. При этом на первый план выходят методы разведочного анализа данных и нелинейного моделирования, в частности нейросетевые технологии прогнозирования [45, 46] свойств сложных систем.

Вычислительные методы разведочного анализа данных включают основные статистические методы (процедура анализа распределений переменных, просмотр корреляционных матриц, анализ многовходовых таблиц частот), а также более сложные, специально разработанные методы анализа, предназначенные для отыскания законо-

мерностей в многомерных данных – факторный анализ, кластерный анализ (древовидная классификация, метод *k*-средних), дискриминантный анализ, канонический анализ, построение деревьев классификации [47].

В работах [1, 48–55] изложены математические основы алгоритмов разведочных, регрессионных и нейросетевых методов анализа, приведена интерпретация статистических дефиниций, показателей, терминов и критериев, в работах [56–59] показано их применение к конкретным катион-краун-эфирным системам и равновесиям диссоциации в жидких средах, а в работах [60–65] обсуждаются комплементарные, химические и сольватационно-термодинамические аспекты катион-краун-эфирного комплексообразования в растворах.

Ч. Педерсеном (1970) было выявлено, что многие макроциклические полиэфиры, содержащие 5–15 атомов кислорода, образуют устойчивые комплексы с солями любого из следующих элементов периодической таблицы Д.И. Менделеева – группы Ia (Li^+, Na^+, K^+, Rb^+, Cs^+), Iб (Ag^+, Au^+), IIa (Ca^{2+}, Sr^{2+}, Ba^{2+}), IIб (Cd^{2+}, Hg^+, Hg^{2+}), IIIa (La^{3+}, Ce^{3+}), IIIб (Tl^+) и IVб (Pb^{2+}).

Краун-эфиры нашли применение во многих областях науки и техники благодаря их способности избирательно распознавать катионы разного заряда и размера [66]: в аналитической химии селективные катионо-связывающие свойства краун-эфиров используются в разделительных и транспортных технологиях для обогащения или извлечения катионов, при конструировании ионоселективных электродов, в хроматографических методах в качестве стационарной фазы; во многих органических синтезах; в качестве катализаторов в межфазном катализе; при имитировании ферментативной активности и разработке новых фармацевтических препаратов; в медицине в качестве диагностических или терапевтических средств. Это далеко неполный перечень практических приложений уникальных комплексообразующих свойств краун-эфиров. В научной практике краун-эфиры и краун-соединения применяются как модели природных ионофоров для исследования механизма транспорта катионов через биологические мембраны. Принципиальным преимуществом синтетических макроциклических ионофоров является хорошо идентифицированная структура, характеризуемая наличием внутримолекулярной полости, для включения катионов в краун-эфир и краун-эфирные фрагменты путем нековалентных взаимодействий [67]: ионных, ион-дипольных, ван-дер-ваальсовых, гидрофобных и водородных связей, формирующих супрамолекулярные структуры.

Целью данной работы является дальнейшее развитие математических моделей прогнозирования [43] термодинамических констант комплексообразования краун-эфиров с катионами по свойствам лиганда, взаимодействующего с ним катиона и используемого растворителя на основе разведочных, регрессионных и нейросетевых алгоритмов анализа данных.

Мерой устойчивости комплексов краун-эфиров с катионами является термодинамическая константа устойчивости $K = [LM^+]\gamma_{LM^+}/[L]\gamma_L[M^+]\gamma_{M^+}$, отвечающая простейшей схеме комплексообразования: $L + M^+ = LM^+$, где [L] и γ_L, $[M^+]$ и γ_{M^+}, $[LM^+]$ и γ_{LM^+} – концентрации и коэффициенты активности свободного лиганда, катиона и комплекса соответственно.

Важность разработки модельного подхода (математических моделей) [43] к анализу, обобщению и прогнозированию устойчивости краун-эфирных комплексов катионов, обусловлена с одной стороны наличием обширного экспериментального материала по термодинамике комплексообразования, а с другой – отсутствием обоснованных критериев выбора оптимального растворителя для управления процессом комплексообразования.

Поэтому актуальным является совместное использование разведочных, регрессионных, нейросетевых алгоритмов и сольватационно-термодинамических подходов [34–37, 43, 56–65] для анализа и прогнозирования термодинамики образования коронатов в воде, неводных и водно-органических растворителях.

Для проведения компьютерного моделирования использованы литературные данные по константам устойчивости комплексов краун-эфиров с катионами из фундаментального обзора [66], в котором собраны результаты кондуктометрического исследования образования коронатов катионов в разных растворителях за четыре десятилетия подряд с 1970 по 2011 г. и представленные в 107 пу-

Таблица 1. Описательная статистика показателей комплексообразования в разных растворителях, отобранных для разведочного анализа

Показатель	Количество значений	Среднее	Минимальное значение	Максимальное значение	Стандартное отклонение	Стандартная ошибка
lg$K$	131	3.71	0.73	7.75	1.34	0.12
M_L	7	386.54	176.20	536.60	126.44	11.05
r_M	15	1.32	0.72	1.70	0.29	0.03
r_L	5	1.80	0.60	3.00	0.83	0.07
d	9	4.70	3.43	5.84	0.53	0.05
B_{KT}	9	0.42	0.06	0.69	0.20	0.02
E_T	9	0.50	0.32	1.00	0.15	0.01
ε	9	37.76	10.36	78.36	14.54	1.27

бликациях в 20 авторитетных научных журналах мира.

Компьютерное моделирование проведено в средах STATISTICA 12 и SPSS 23 на платформе Windows 10 для комплексов состава 1:1 краун-эфиров (12C4, 16C5, $(CH_3)_2$16C5, DB21C7, DB24C8, DCH24C8, DB30C10) с катионами щелочных (Li^+, Na^+, K^+, Cs^+, Rb^+), щелочноземельных (Mg^{2+}, Ca^{2+}, Sr^{2+}, Ba^{2+}), тяжелых (Ag^+, Tl^+, Co^{2+}, Cu^{2+}, Pb^{2+}) металлов и NH_4^+ в воде (W) и органических растворителях (метаноле, ацетонитриле, ацетоне, *N*,*N*-диметилформамиде, нитробензоле, нитрометане, 1,2-дихлорэтане, пропиленкарбонате) при 298.15 K.

Построены корреляционные матрицы свойств растворителей, катионов и краун-эфиров. Методом главных компонент [47, 50, 51, 54] отобраны для построения математических моделей параметры (свойства): растворителей – диэлектрическая проницаемость ε [68], параметры Димрота–Райхардта E_T и Камлета–Тафта B_{KT} [68], диаметр молекулы растворителя d, Å [68]; катионов – эффективный ионный кристаллохимический радиус для координационного числа 6 r_M, Å [69]; краун-эфиров – молекулярная масса M_L и радиус полости краун-эфира r_L [66, 70, 71].

Поставленная цель достигнута путем решения следующих задач: (1) первичный анализ данных, вычисление описательных статистик, проверка нормальности распределения; (2) факторный анализ – построение корреляционных матриц, выделение латентных факторов; (3) кластерный анализ – алгоритм древовидной кластеризации, итерационный алгоритм *k*-средних; (4) дискриминантный анализ Фишера – построение линейных классификационных функций; (5) канонический дискриминантный анализ – построение канонических линейных дискриминантных функций; (6) деревья классификации – построение дендрограммы и правила кластеризации устойчивости коронатов; (7) регрессионный анализ зависимости устойчивости коронатов от свойств среды, катионов и краун-эфиров; (8) нейросетевой анализ – нейросетевой классификатор, нейросетевой аппроксиматор; (9) прогностические возможности регрессионных и нейросетевых моделей.

Первичный анализ данных. В табл. 1 приведены количественные параметры описательной статистики [72, 73] отобранных для анализа показателей. Среднее квадратическое отклонение (стандартное отклонение) данных меньше половины среднего арифметического, поэтому распределение можно считать симметричным. Проверка гипотезы нормального распределения анализируемых данных (табл. 2) выполнена по критериям Шапиро–Уилка ($8 < n < 50$) и Колмогорова–Смирнова ($n > 50$) [73, 74].

Факторный анализ. Надежность вычислений элементов корреляционной матрицы и целесообразность ее описания с помощью факторного анализа [49–51] подтверждены мерой адекватности выборки Кайзера–Мейера–Олкина (критерий

Таблица 2. Расчетные и табличные (критические) значения критериев проверки гипотезы нормальности распределения переменных[a]

Переменная, (n)	Критерий Шапиро–Уилка, $W_{расч}$ ($W_{табл}$)	Критерий Колмогорова–Смирнова, $D_{расч}$ ($D_{табл}$)
lg$K$, (131)		0.075 (0.119)
r_M, (15)	0.888 (0.881)	
d, (9)	0.920 (0.829)	
B_{KT}, (9)	0.842 (0.829)	
E_T, (9)	0.748 (0.829)	
ε, (9)	0.776 (0.829)	

[a] n – объем выборки, p – уровень значимости. Если табличное значение $W_{табл}$ меньше расчетного значения $W_{расч}$, а $D_{табл} > D_{расч}$, то распределение считается соответствующим нормальному на уровне значимости $p = 0.05$.

Таблица 3. Корреляционная матрица показателей равновесия комплексообразования

Показатели	Коэффициенты корреляции							
	lg$K$	M_L	r_M	r_L	d	B_{KT}	E_T	DP
lg$K$	1.00	0.51	0.01	0.50	0.22	–0.74	–0.43	–0.29
M_L	0.51	1.00	0.06	0.98	0.24	–0.30	–0.40	–0.35
r_M	0.01	0.06	1.00	0.06	0.03	–0.08	–0.06	0.01
r_L	0.50	0.98	0.06	1.00	0.23	–0.32	–0.41	–0.39
d	0.22	0.24	0.03	0.23	1.00	–0.32	–0.66	0.21
B_{KT}	–0.74	–0.30	–0.08	–0.32	–0.32	1.00	0.39	0.08
E_T	–0.43	–0.40	–0.06	–0.41	–0.66	0.39	1.00	0.30
ε	–0.29	–0.35	0.01	–0.39	0.21	0.08	0.30	1.00

Таблица 4. Факторные нагрузки, собственные значения и веса факторов[a]

Переменные	Факторные нагрузки		
	Фактор 1 (F_1)	Фактор 2 (F_2)	Фактор 3 (F_3)
lg$K$	–0.623	0.486	–0.074
M_L	–0.871	0.199	0.089
r_M	–0.042	0.018	0.987
r_L	–0.885	0.194	0.078
d	0.033	0.887	0.043
B_{KT}	0.357	–0.635	–0.015
E_T	0.384	–0.712	–0.007
ε	0.719	0.269	0.112
Собственные значения	2.724	2.082	1.009
Вес фактора, %	0.340	0.260	0.126

[a] Фактор – латентная (скрытая) переменная, конструируемая таким образом, чтобы можно было объяснить корреляцию между набором переменных; факторные нагрузки – линейные корреляции между переменными и факторами; собственное значение – представляет полную дисперсию, объясняемую каждым фактором; вес фактора – процент от полной дисперсии, приписываемый каждому фактору [50].

KMO = 0.573 – коэффициент, для проверки целесообразности выполнения факторного анализа) и коэффициентом сферичности Бартлетта (критерий Хи-квадрат = 735.92, значимость критерия Бартлетта $p < 0.001$) [50]. Высокие значения KMO (от 0.5 до 1) указывают на целесообразность факторного анализа данных. Критерий Бартлетта – статистика, проверяющая гипотезу о том, что переменные в генеральной совокупности не коррелируют между собой, если p-уровень не превышает 0.05.

Методом главных компонент [47, 49–51] по выборочной совокупности значений семи отобранных показателей вычислены корреляционная матрица системы используемых для анализа данных (табл. 3), ее собственные значения, факторные нагрузки и веса факторов (табл. 4) [50]. Свойства растворителей и краун-эфиров проявляют как умеренную положительную (M_L, r_L, d), так и отрицательную (B_{KT}, E_T, ε) взаимосвязь с lg$K$, радиус катионов (r_M) демонстрирует слабую зависимость

Таблица 5. Коэффициенты факторных моделей

Праметр	Коэффициенты a_i факторных моделей		
	F_1	F_2	F_3
lg$K$, a_1	–0.177	0.159	–0.100
M_L, a_2	–0.338	–0.058	0.071
r_M, a_3	–0.004	–0.039	0.982
r_L, a_4	–0.345	–0.063	0.060
d, a_5	0.185	0.508	0.006
B_{KT}, a_6	0.032	–0.291	0.015
E_T, a_7	0.029	–0.330	0.027
ε, a_8	0.364	0.286	0.108

Таблица 6. Результаты дисперсионного анализа стандартизированных показателей комплексообразования методом k-средних[a]

Показатель	SS_B	df_B	SS_W	df_W	$F(4, 126)$	p
lgK_{st}	73.29	4	56.71	126	40.71	0.000
$M_{L,st}$	94.11	4	35.89	126	82.59	0.000
$r_{M,st}$	8.18	4	121.82	126	2.12	0.083
$r_{L,st}$	85.89	4	44.11	126	61.33	0.000
d_{st}	62.67	4	67.33	126	29.32	0.000
$B_{KT.st}$	108.67	4	21.33	126	160.51	0.000
$E_{T,st}$	112.84	4	17.16	126	207.17	0.000
ε_{st}	42.53	4	87.47	126	15.32	0.000

[a] SS_W – сумма квадратов отклонений значений каждого из предикторов (свойство растворителя, краун-эфира и катиона) от группового среднего значения предиктора внутри группы (кластера) – мера внутригрупповой изменчивости: $\sigma^2_{SS_W} = SS_W/(n-1)$, где $\sigma^2_{SS_W}$ – внутригрупповая дисперсия; SS_B – межгрупповая сумма квадратов отклонений средних значений предикторов в каждой из групп от суммарного среднего значения предикторов по всем группам – мера межгрупповой изменчивости: $\sigma^2_{SS_B} = SS_B/(n-1)$, где $\sigma^2_{SS_B}$ – межгрупповая дисперсия; значение критерия Фишера $F = MS_B/MS_W$, где $MS_B = SS_B/df_B$, $MS_W = SS_W/df_W$; MS_W и MS_B – средние значения квадратов отклонений внутри групп и между ними; $df_W = (n - m - 1)$ и $df_B = (m - 1)$ – соответствующие степени свободы (m – число групп, n – количество наблюдений в каждой из групп); $F(4, 126)$ – наблюдаемый критерий Фишера. [$F_{кр}(4, 126, p = 0.05) = 2.44$]; p – наблюдаемый уровень значимости [49–51].

от константы комплексообразования, свойств краун-эфиров и растворителей (табл. 3).

Нагрузки латентных факторов (F_1, F_2 и F_3) определены методом главных компонент с использованием критерия каменистой осыпи и процедуры ортогонального варимакс-вращения факторов [50]. Метод главных компонент – один из основных способов уменьшить размерность данных, потеряв наименьшее количество информации. Процедура ортогонального варимакс-вращения – метод вращения факторов, минимизирующий число переменных с высокими нагрузками на каждый фактор. Критерий каменистой осыпи (Cattell, 1966) состоит в поиске точки на графике зависимости собственных значений от числа факторов, где убывание собственных значений замедляется наиболее сильно.

Для анализа отобрано три фактора, собственные значения которых больше единицы. Первый фактор объясняет 34.0% суммарной дисперсии, второй фактор – 26% и третий фактор – 12.6% (табл. 4).

Переменные M_L и r_L коррелируют с фактором 1, коэффициент корреляции равен –0.871 и –0.885 соответственно, переменные d и E_T коррелируют с фактором 2 (0.887 и –0.712) и переменная r_M нагружена третьим фактором (0.987).

Таким образом, преимущественно первый фактор связан с вариациями свойств краун-эфиров, второй фактор – с изменением свойств растворителей и третий фактор зависит от радиуса катионов.

В табл. 5 приведены коэффициенты уравнения a_i (1) для трех факторных моделей, полученных методом главных компонент с применением варимакс-вращения факторов [50].

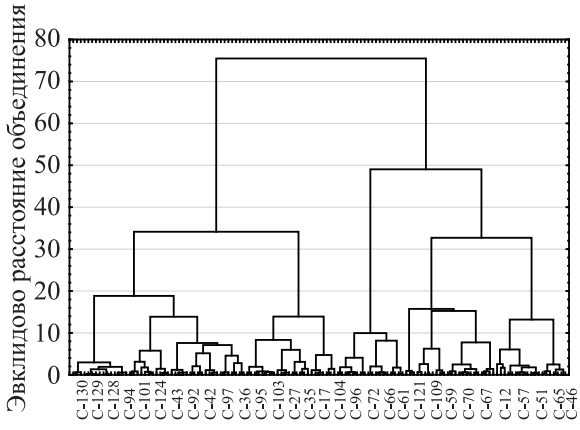

Рис. 1. Дендрограмма иерархической кластеризации констант устойчивости коронатов.

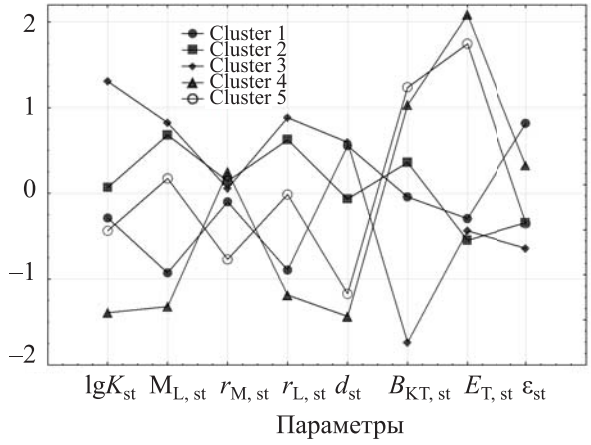

Рис. 2. Средние значения показателей комплексообразования для пяти групп констант устойчивости коронатов катионов.

$$F_j = a_1\lg K + a_2 M_L + a_3 r_M + a_4 r_L + a_5 d + a_6 B_{KT} + a_7 E_T + a_8\varepsilon, j = 1\text{–}3. \quad (1)$$

Анализ рассчитанных факторов F_1, F_2 и F_3 позволяет выяснить, какие эффекты превалируют в устойчивости коронатов в растворе – эффекты среды, свойства катионов или краун-эфиров по максимальному значению фактора.

Кластерный анализ. В работе реализованы два метода кластерного анализа [49, 50], представленные в статистическом пакете STATISTICA 12 [47, 51, 54]: агломеративный – объединение, или дерево кластеризации и дивизивный – кластеризация *k*-средними. Предварительно была проведена процедура стандартизации исходных данных (z-оценки) путем вычитания среднего и деления на стандартное отклонение.

Агломеративная кластеризация. На рис. 1 приведена дендрограмма иерархической кластеризации устойчивости 131 короната по свойствам растворителей, катионов и краун-эфиров.

Объединение констант устойчивости коронатов в кластеры проведено методом Варда [49, 50, 51] с использованием Евклидового расстояния в качестве метрики пространства. В отличие от всех других методов в методе Варда используется алгоритм дисперсионного анализа для оценки расстояний между кластерами. Евклидово расстояние является геометрическим расстоянием в многомерном пространстве и вычисляется по формуле (2). Расстояние между точками:

$$(x, y) = \{\sum_i (x_i - y_i)^2\}^{1/2}. \quad (2)$$

На евклидовом расстоянии, равном 20, выявлено 5 кластеров; при увеличении расстояния до 40 количество кластеров равно трем, на расстоянии 60 – 2 кластера.

Кластерный анализ алгоритмом k-средних. Наилучшее согласие результатов двух методов кластерного анализа получено при выборе 5 кластеров. На рис. 2 приведен график средних значений показателей комплексообразования для пяти кластеров, отображающих различие между группами констант устойчивости коронатов по каждому из свойств.

Результаты дисперсионного анализа свидетельствуют (табл. 6), что распределение констант устойчивости по кластерам проведено удовлетворительно. Уровень значимости p у критерия Фишера значительно меньше 0.05 для всех переменных, а наблюдаемый критерий Фишера больше критического $F_{набл} > F_{кр}$, за исключением $r_{M,st}$.

Количественный (131 константа) состав кластеров: первый кластер объединяет 37 констант устойчивости коронатов в апротонных растворителях: MeCN, ацетоне, ДМФА, пропиленкарбонате, нитробензоле; второй кластер – 48 констант устойчивости коронатов в апротонных растворителях: MeCN, ацетоне, ДМФА; третий кластер группирует 24 константы устойчивости коронатов в 1,2-дихлорэтане и нитробензоле; четвертый кластер – 14 констант устойчивости коронатов в

Таблица 7. Результаты дискриминантного анализа (алгоритм – переменные в модели)

Свойство	Группирующая переменная: 5 кластеров констант устойчивости коронатов; Λ-Уилкса: 0.0001; $F_{набл}(28, 434) = 221.43$, $p < 0.000$; $F_{кр}(28, 434) = 1.03$, $F_{кр}(4, 120) = 2.45$					
	Λ-Уилкса	Частная Λ-Уилкса	$F_{искл}(4, 120)$	p-уровень	Толерантность, 1-R^2	R^2
M_L	0.0001	0.576	22.10	0.000	0.099	0.901
r_M	0.0001	0.917	2.70	0.034	0.966	0.034
r_L	0.0001	0.791	7.94	0.000	0.105	0.895
d	0.0011	0.050	570.93	0.000	0.035	0.965
B_{KT}	0.0005	0.107	251.35	0.000	0.519	0.481
E_T	0.0082	0.007	4559.35	0.000	0.017	0.983
ε	0.0023	0.023	1280.81	0.000	0.018	0.982

[a] R^2 – коэффициент множественной корреляции данного свойства со всеми остальными свойствами, использованными в анализе.

Таблица 8. Коэффициенты линейных классификационных функций Фишера[a]

Параметр, b_i	Кластер 1 $p = 0.282$	Кластер 2 $p = 0.351$	Кластер 3 $p = 0.183$	Кластер 4 $p = 0.107$	Кластер 5 $p = 0.076$
M_L, b_1	0.03	0.18	0.11	0.02	0.2
r_M, b_2	–13.14	–13.18	–19.33	–20.13	–25.7
r_L, b_3	58.07	47.88	60.86	87.63	71.7
d, b_4	1382.94	1389.21	1556.93	1893.35	1908.1
B_{KT}, b_5	–268.53	–235.99	–397.27	–264.56	–260.0
E_T, b_6	14098.31	14263.48	15782.90	19673.96	19851.5
ε, b_7	–61.39	–62.22	–69.20	–85.35	–86.3
b_8	–5124.69	–5242.44	–6438.49	–9860.81	–10052.6

[a] p – апостериорные (послеопытные) вероятности [50, 73], пропорциональные числу констант комплексообразования катионов с краун-эфирами в кластере.

протолитических растворителях (вода и MeOH) и пятый кластер содержит 10 констант в MeOH.

Распределение семи краун-эфиров по кластерам: 1, 2, 3, 5 кластеры – DB24C8; 1, 2, 3 кластеры – DB21C7; 1, 4, 5 – 16C5, $(CH_3)_2$16C5; 2, 3 кластеры – DB30C10; 1, 4 кластеры – 12C4; 2 кластер – DCH21C7.

Распределение 15 катионов по кластерам: 1 – 5 кластеры – катионы щелочных металлов (Na^+, K^+, Rb^+, Cs^+); 1 – 4 кластеры – Tl^+, 1 – 3 кластеры NH_4^+; 1, 3, 4 кластеры – Li^+; 2, 5 кластеры – Mg^{2+}, Ca^{2+}; 2, 4 кластеры – Sr^{2+}, Ba^{2+}; 2, 3 кластеры – Ag^+; 5 кластер – Co^{2+}, Cu^{2+}, Pb^{2+}.

Всегда следует иметь в виду приближенный характер моделей. Ни один отдельный статистический анализ не является универсальным и достаточным для установления степени научной обоснованности полученных результатов. Для этого требуется подтверждение результатов моделирования на других наборах данных и разными алгоритмами, исключающими появление систематических ошибок из-за неправильного использования данных [75]. Поэтому для подтверждения результатов иерархического кластерного анализа (агломеративная кластеризация, метод k-средних) разработаны дискриминантная [47, 49–51], каноническая [47, 49–51], дерево решений [53, 76] и нейросетевые [52, 53, 55] модели классификации и аппроксимации (регрессии) констант устойчивости коронатов катионов.

Дискриминантный анализ. Для разделения констант устойчивости коронатов на группы по свойствам растворителей, краун-эфиров и катионов проведен линейный дискриминантный анализ Фишера, реализованный в статистическом пакете STATISTICA 12 [47, 51, 54].

Результаты, полученные при одновременном введении всех переменных в дискриминантный анализ, даны в табл. 7. Λ-Уилкса для каждого предиктора – это отношение внутригрупповой суммы квадратов отклонений предиктора от выборочного среднего к общей сумме квадратов отклонений, иначе говоря – это отношение меры внутригрупповой изменчивости SS_W к мере общей изменчивости

Таблица 9. Матрица кластеризации констант устойчивости коронатов[a]

Кластер	Доля правильной кластеризации, %	Кластер 1 $p = 0.282$	Кластер 2 $p = 0.351$	Кластер 3 $p = 0.183$	Кластер 4 $p = 0.107$	Кластер 5 $p = 0.076$
Кластер 1	100.0	37	0	0	0	0
Кластер 2	100.0	0	46	0	0	0
Кластер 3	100.0	0	0	24	0	0
Кластер 4	100.0	0	0	0	14	0
Кластер 5	80.0	0	0	0	2	8
Всего, %	98.5	37	46	24	16	8

[a] Строки матрицы – наблюдаемая кластеризация методом *k*-средних. Столбцы матрицы – предсказанная классификация дискриминантным анализом Фишера.

Таблица 10. Характеристика извлеченных канонических корней (канонических линейных дискриминантных функций)[a]

Извлечено корней	Хи-квадрат – критерий последовательности удаления корней						
	Co	R	R^2	Λ	χ^2	ν	p
0	235.60	0.9979	0.9958	0.0001	1219.10	28	0.000
1	12.23	0.9615	0.9245	0.0127	541.27	18	0.000
2	4.14	0.8976	0.8057	0.1682	221.07	10	0.000
3	0.16	0.3674	0.1350	0.8650	17.98	4	0.001

[a] Co – собственное значение, R – коэффициент канонической корреляции, R^2 – коэффициент детерминации, Λ – значение статистики Λ-Уилкса, χ^2 – значение статистики Хи-квадрат Пирсона, ν – число степеней свободы, p – уровень значимости соответствующего канонического корня.

$SS_{Total} = SS_W + SS_B$. Значение стандартной статистики Уилкса лямбда (Λ-Уилкса) равно 0.0001, что свидетельствует о высокой дискриминирующей мощности модели (1.0 – дискриминация отсутствует, 0.0 – полная дискриминация). Этот вывод также подтверждается наблюдаемым значением $F_{набл}$-статистики, $F_{набл}(28, 434) = 221.43$, $p < 0.000$ и $F_{набл}(28, 434) > F_{кр}(28, 434)$.

Из анализа результатов табл. 7 следует, что только три свойства – E_T, ε и d – демонстрируют способность к дискриминации, чем больше значение Λ-Уилкса, тем более предпочтительным является это свойство в процедуре разделения констант устойчивости коронатов по группам.

Частная Λ-Уилкса, характеризующая единичный вклад соответствующей переменной в разделительную силу модели, подтверждает этот вывод. Она равна отношению Λ-Уилкса после добавления переменной в модель к Λ-Уилкса до добавления этой переменной [73]. Чем меньше значение частной Λ-Уилкса, тем больший вклад этого свойства в общую дискриминацию. Наряду с этим, чем меньше значение критерия Фишера $F_{искл}$ (табл. 7), тем менее желательны свойства в модели дискриминации. Переменные, у которых уровень значимости $p > 0.05$, исключаются из дискриминантной модели. Толерантность является мерой избыточности переменной в модели (чем меньше ее значение, тем избыточнее переменная в модели, тем меньшую дополнительную информацию несет эта переменная (свойство) [47], иначе говоря, чем ниже толерантность, тем сильнее данное свойство связано (коррелирует) со всеми остальными (наличие мультиколлиниарности).

В табл. 8 приведены коэффициенты b_i математической модели дискриминации констант устойчивости коронатов – линейных классификационных функций (ЛКФ).

$$\text{Кластер } J = b_1 M_L + b_2 r_M + b_3 r_L + b_4 d + b_5 B_{KT} + b_6 E_T + b_7 \varepsilon + b_8, j = 1\text{–}5. \quad (3)$$

Подставив в эти уравнения значения свойств растворителя, катиона и краун-эфира, которые не использовались при построении линейных классификационных функций, можно предсказать кластер, к которому константа устойчивости будет отнесена по наибольшему рассчитанному значению линейной классификационной функции [36].

В табл. 9 представлен конечный результат дискриминантного анализа – матрица кластеризации [50]. На диагонали матрицы содержится количество констант устойчивости коронатов правильно

Таблица 11. Коэффициенты канонических линейных классификационных функций Фишера

Параметр, A_i	DF_1	DF_2	DF_3	DF_4
M_L, A_1	0.0003	0.0016	–0.0329	0.0226
r_M, A_2	0.2376	0.4307	0.2722	–2.4149
r_L, A_3	–0.7031	–0.2718	2.7175	–3.5982
d, A_4	–12.8920	–4.3274	1.2789	1.2964
B_{KT}, A_5	0.4069	16.5575	–2.7895	0.3113
E_T, A_6	–139.5100	–16.9536	1.6234	2.2455
ε, A_7	0.6005	0.1270	0.0353	–0.0108
A_8	108.5070	16.4554	0.4617	–6.0258

Таблица 12. Средние канонических переменных (центроиды кластеров)

Кластер	DF_1	DF_2	DF_3	DF_4
Кластер 1	9.977	0.852	2.568	0.240
Кластер 2	8.583	2.096	–1.959	–0.167
Кластер 3	–1.165	–7.160	–0.499	–0.068
Кластер 4	–30.149	1.994	1.964	–0.675
Кластер 5	–31.393	1.596	–2.040	0.987

классифицированных в кластеры. В пятый кластер правильно отнесены 8 констант устойчивости коронатов из 10 (80.0% правильной кластеризации). Две константы устойчивости ошибочно отнесена к четвертому кластеру. Дискриминантный анализ был выполнен в трех режимах, представленных в пакете STATISTICA 12: в стандартном (Standard), пошаговом вперед (Forward stepwise) и пошаговом назад (Backward stepwise) [51]. При этом дискриминантная модель кластеризации констант устойчивости коронатов на 98.5% подтвердила результаты метода *k*-средних.

Канонический анализ. Для получения дополнительных сведений о природе дискриминации (разделения) констант устойчивости проведен канонический анализ [47, 50]**.** Показано как семь переменных (свойства растворителей, краун-эфиров и катионов) разделяют константы устойчивости коронатов каноническими линейными дискриминантными функциями (КЛДФ) на 5 групп, выделенных методом *k*-средних.

Извлечены четыре независимые (ортогональные) дискриминирующие функции (табл. 10). Первая строка содержит критерий значимости для всех дискриминантных функций (корней). Так как уровень значимости p меньше 0.05, то имеется хотя бы один канонический корень, который является статистически значимым, вторая строка характеризует значимость дискриминантных функций, оставшихся после удаления первой. Так как $p < 0.05$, среди оставшихся корней есть статистически значимые, в третьей строке содержатся данные о значимости функций, оставшихся после удаления первых двух ($p < 0.05$). Каждая последующая дискриминантная функция вносит все меньший и меньший вклад в общую дискриминацию. Из анализа данных табл. 10 вытекает, что все извлеченные корни (дискриминантные функции) статистически значимы, так как уровень значимости p меньше 0.05 [47]. Собственное (характеристическое) значение для каждой дискриминантной функции – это отношение межгрупповой суммы квадратов отклонений SS_B к внутригрупповой сумме квадратов отклонений SS_W. Большие собственные значения свидетельствуют о высокой статистической значимости извлеченных дискриминантных корней (функций).

Чем больше теоретические числа, рассчитанные на основе нулевой гипотезы (отсутствие различий между кластерами), будут отличаться от фактических, тем сильнее критерий Хи-квадрат будет отличаться от 0 (Λ-Уилкса, наоборот, будет приближаться к 0), тем с большей вероятностью можно принять альтернативную статистическую гипотезу и говорить о статистической достоверности имеющихся различий в сравниваемых кластерах.

В табл. 11 приведены коэффициенты A_i (i = 1–8) канонических моделей – канонических линейных дискриминантных функций DF_j (j = 1–4), для ис-

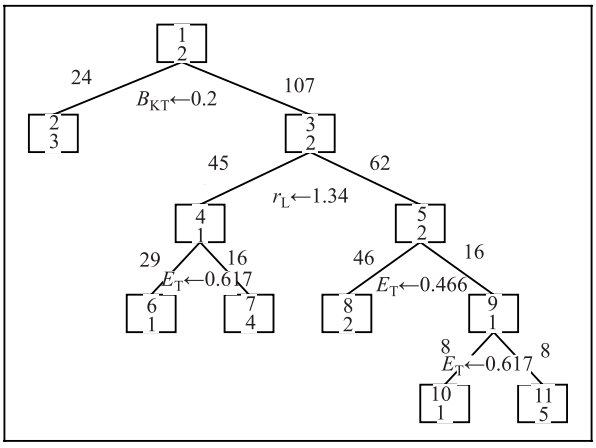

Рис. 3. Граф дерева классификации устойчивости коронатов катионов.

ходных (нестандартизированных) свойств катионов, краун-эфиров и растворителей.

$$DF_j = A_1M_L + A_2r_M + A_3r_L + A_4a + A_5B_{KT} + A_6E_T + A_7\varepsilon + A_8. \quad (4)$$

Константа устойчивости исследуемого коронаата катиона, для которой по свойствам растворителя, краун-эфира и катиона рассчитаны канонические линейные дискриминантные функции DF_1, DF_2, DF_3 и DF_4, будет отнесена к кластеру по наименьшему расстоянию до центра (центроида) соответствующего кластера [36]. В табл. 12 приведены координаты центроидов четырех кластеров констант устойчивости коронатов.Таблица средних значений для дискриминантных функций (табл. 12) позволяет определить кластеры, лучше всего идентифицируемые конкретной дискриминантной функцией. Функция DF_1 идентифицирует в основном кластеры 4 и 5, так как им соответствуют наибольшие значения этой функции. Функция DF_2 – кластеры 2 и 3. Функция DF_3 – кластер 1. Функция DF_4 – кластер 5.

Деревья классификации [53, 76]**.** Методологические аспекты построения деревьев классификации (правил решения) констант диссоциации и комплексообразования алгоритмом CART (Classification and Regression Trees) изложены в работах [34, 36]. На рис. 3 приведен граф дерева классификации устойчивости коронатов катионов – четыре вершины ветвления (1, 3, 4, 5, 9) и шесть терминальных вершин (2, 6, 7, 8, 10, 11) – обозначения в верхней части вершин. Текст под вершинами ветвления описывает условие ветвления. Числа в нижней части вершин обозначают номер кластера. Числа над вершинами показывают количество констант устойчивости коронатов, отнесенных к данной вершине. Все константы устойчивости коронатов в вершинах ветвления относятся к кластеру, в котором количество констант устойчивости наибольшее. Поэтому корневая вершина ветвления 1 обозначена как Кластер 2.

В табл. 13 приведена структура дерева классификации устойчивости коронатов катионов по свойствам растворителей (d, E_T, B_{KT}, ε), катионов

Таблица 13. Структура дерева классификации устойчивости коронатов

Вершина	Левая вершина	Правая вершина	Кластер 1	Кластер 2	Кластер 3	Кластер 4	Кластер 5	Предсказанный кластер	Константа ветвления	Переменная ветвления
1	2	3	37	46	24	14	10	2	0.200	B_{KT}
2			0	0	24	0	0	3		–
3	4	5	37	46	0	14	10	2	1.340	r_L
4	6	7	29	0	0	14	2	1	0.617	E_T
5	8	9	8	46	0	0	8	2	0.466	E_T
6			29	0	0	0	0	1		–
7			0	0	0	14	2	4		–
8			0	46	0	0	0	2		–
9	10	11	8	0	0	0	8	1	0.617	E_T
10			8	0	0	0	0	1		–
11			0	0	0	0	8	5		–

Таблица 14. Итоги нейросетевого аппроксиматора MLP 7-7-1[a]

Архитектура	Производительность обучения	Контрольная производительность	Тестовая производительность	Ошибка обучения	Контрольная ошибка	Тестовая ошибка	Алгоритм обучения	Функция ошибки	Функция активации скрытых нейронов	Функция активации выходных нейронов
MLP 7-7-1	0.958	0.909	0.902	0.08	0.15	0.15	BFGS 116	SOS	Logistic	Exponent

[a] Производительность обучения, контрольная производительность, тестовая производительность – отношение стандартного отклонения ошибки прогноза к стандартному отклонению исходных данных на соответствующих выборках; Ошибка обучения, контрольная ошибка, тестовая ошибка – ошибки сети на соответствующих выборках; BFGS – алгоритм Бройдена–Флетчера–Гольдфарба–Шанно [77, 78]; SOS – среднеквадратичная ошибка $E = \frac{1}{P}\sum_{i=1}^{P}(\lg K_{\text{расч},i} - \lg K_{\text{эксп},i})^2$, P – количество обработанных примеров в выборке; Exponent – экспоненциальная функция $\varphi(x) = e^x$; Logistic – логистическая функция $\varphi(x) = 1/[1 + \exp(-tx)]$.

(r_M) и краун-эфиров (M_L, r_L). Ранги значимости предикторов дерева кластеризации d, B_{KT}, E_T, ε, r_M, M_L, r_L равны 88, 80, 100, 72, 12, 77, 76 соответственно (0 – низкая значимость, 100 – высокая значимость).

Проведенный кластер-анализ с использованием деревьев решений на 98.5% подтвердил (рис. 3, табл. 13) результаты кластерного анализа устойчивости коронатов методом *k*-средних. Как и в случае дискриминантного анализа, составы первого, второго, третьего и четвертого кластеров подтверждены на 100%. Две константы устойчивости пятого кластера алгоритмом CART ошибочно отнесены в четвертый кластер (80.0%).

Регрессионный анализ. Математические регрессионные модели [51, 53, 54] имеют вид:

– включены все переменные (standard)

$$\lg K = (6.52 \pm 2.55) + (0.01 \pm 0.00)M_L - (0.30 \pm 0.47)r_M - (0.92 \pm 0.76)r_L - (0.27 \pm 0.43)d - (4.46 \pm 0.76)B_{KT} - (1.14 \pm 1.55)E_T - (0.01 \pm 0.02)\varepsilon. \quad (5)$$

R = 0.8252, наблюдаемое значение критерия Фишера $F_{\text{набл}}(7,123)$ = 37.51, p < 0.000, критическое значение критерия Фишера $F_{\text{кр}}(7,123)$ = 2.01, p = 0.05, стандартная ошибка = 0.78, критерий Дарбина–Уотсона d_{DW} = 1.38.

– отбор переменных методом прямого выбора (forward selection)

$$\lg K = (4.85 \pm 1.06) + (0.01 \pm 0.00)M_L - (0.28 \pm 0.47)r_M - (0.94 \pm 0.76)r_L - (4.53 \pm 0.73)B_{KT} - (0.02 \pm 0.02)\varepsilon,$$
$$R = 0.8218, F_{\text{набл}}(5,125) = 51.99,$$
$$p < 0.000, F_{\text{кр}}(5,125) = 2.21, p = 0.05, \quad (6)$$

стандартная ошибка = 0.78, d_{DW} = 1.32.

– отбор переменных методом обратного исключения (backward elimination)

$$\lg K = (4.20 \pm 0.63) - (4.34 \pm 0.74)B_{KT} - (0.0034 \pm 0.0012)M_L,$$
$$R = 0.8000, F_{\text{набл}}(2,128) = 190.29,$$
$$p < 0.000, F_{\text{кр}}(2,128) = 3.00, p = 0.05, \quad (7)$$

стандартная ошибка = 0.81, d_{DW} = 1.46.

Критерий Дарбина-Уотсона (d_{DW}) [51] применяется при анализе остатков регрессионных моделей для тестирования автокорреляции первого порядка переменных исследуемых моделей. Автокорреляция остатков наблюдается тогда, когда значения предыдущих остатков завышают (положительная) или занижают (отрицательная) значения последующих. Если $0 < d_{DW} < 1.5$ имеется положительная автокорреляция. Одной из причин автокорреляции может быть неучет в регрессионной модели одного или нескольких важных параметров (свойств).

Таблица 15. Итоги кластеризации констант устойчивости коронатов многослойным персептроном MLP 7-5-5

Архитектура	Показатели кластеризации	Кластер 1	Кластер 2	Кластер 3	Кластер 4	Кластер 5	Все
MLP 7-5-5	Все	37	46	24	14	10	131
	Правильно	37	46	24	14	10	131

Таблица 16. Наблюдаемые ($\lg K_{эксп}$) и аппроксимированные ($\lg K_{MLP}$) персептроном MLP 7-1-1 значения констант комплексообразования ($\lg K$) катионов с краун-эфирами

Краун-эфир	Катион	Растворитель	$\lg K_{эксп}$	$\lg K_{MLP}$	Краун-эфир	Катион	Растворитель	$\lg K_{эксп}$	$\lg K_{MLP}$
12C4	Mg^{2+}	Пропиленкарбонат	2.61	1.27	DB18C6	Ag^{+}	Вода	1.41	1.86
15C5	Ag^{+}	Вода	0.94	1.53	DB18C6	Tl^{+}	Вода	1.5	1.75
15C5	Tl^{+}	Вода	1.23	1.08	DB18C6	Pb^{2+}	Вода	1.89	1.93
15C5	Pb^{2+}	Вода	1.85	1.56	DB18C6	Ag^{+}	MeOH	4.04	3.93
18C6	Ag^{+}	Вода	1.55	1.90	DB18C6	Tl^{+}	MeCN	4.90	4.95
18C6	Tl^{+}	Вода	2.27	2.10	DB21C7	Tl^{+}	MeCN	> 5.0	5.36
18C6	Pb^{2+}	Вода	4.3	2.00	DB21C7	Tl^{+}	Ацетон	4.71	4.83
18C6	Ag^{+}	MeOH	4.57	3.96	DB21C7	Tl^{+}	MeOH	3.97	4.38

Таблица 17. Предсказанные моделью MLP 7-1-1 значения констант устойчивости $\lg K_{MLP}$

Катион	Растворитель	Краун-эфир	$\lg K_{эксп}$	$\lg K_{MLP}$	Остатки
Na^{+}	MeOH	24C8	2.35	2.65	–0.3
K^{+}	MeOH	24C8	3.50	3.65	–0.2
Cs^{+}	MeOH	24C8	4.15	3.98	0.2
Ca^{2+}	MeOH	24C8	2.66	2.64	0.0
Tl^{+}	MeOH	DB30C10	4.47	4.89	–0.4
Na^{+}	MeOH	DB30C10	2.10	3.51	–1.4
K^{+}	MeOH	DB30C10	4.60	5.08	–0.5
Rb^{+}	MeOH	DB30C10	4.60	4.92	–0.3
Cs^{+}	MeOH	DB30C10	4.18	4.38	–0.2

Выбранные входные независимые переменные были применены для построения прогностических нейросетевых [45, 55] моделей зависимости констант устойчивости коронатов катионов $\lg K$ от свойств растворителей, катионов и краун-эфиров.

Нейросетевой анализ. В табл. 14 приведены основные характеристики обученного нейросетевого аппроксиматора – многослойного персептрона MLP 7-7-1.

Коэффициенты корреляции на обучающей (70%), контрольной (15%) и тестовой (15%) выборках равны 0.958, 0.909 и 0.902 соответственно. Статистические характеристики обученной нейросетевой модели персептронного типа MLP 7-7-1 (табл. 14) отражают успешность проведенного обучения. Так, качество обучения на различных выборках больше 90%, ошибка обучения на обучающей выборке 0.08, а на контрольной и тестовой выборках 0.15. Эти данные также свидетельствуют о том, что нейросетевая модель обладает большей прогнозирующей силой, чем модели множественной линейной регрессии, коэффициенты корреляции которых меньше 0.83.

Обученный нейросетевой классификатор MLP 7-5-5 (табл. 15) имеет следующие основные характеристики: производительность обучения – 100%, контрольная производительность – 100%, тестовая производительность – 100%; алгоритм обучения – BFGS 57; функция ошибки – SOS; функции активации нейронов: скрытых – логистическая, выходных – тождественная.

Таким образом, алгоритм многослойного персептрона MLP 7-5-5 на 100% подтвердили правомочность кластеризации методом *k*-средних (табл. 15).

Воспроизведение и прогнозирование результатов в новых данных и новых условиях является более надежным способом проверки [75] эффективности построенных нейросетевых моделей (персептронов). Поэтому табл. 16 и 17 в качестве примеров демонстрируют возможности обученного многослойного персептрона MLP 7-7-1 для аппроксимации (табл. 16) и прогнозирования (табл. 17) констант устойчивости коронатов $\lg K_{MLP}$ по свойствам растворителей, катионов и краун-эфиров. При этом важно отметить, что экспериментальные константы комплексообразования $\lg K_{эксп}$ (табл. 17), взятые из работы [78], не использовались в обучении нейронной сети.

На основании анализа данных, приведенных в табл. 16, можно заключить, что обученный нейросетевой аппроксиматор обладает удовлетворительным прогностическим потенциалом, а последующие усовершенствования модели лежат в плоскости пополнения массивов данных как по константам устойчивости комплексов краун-эфиров с катионами, так и по свойствам катионов, краун-эфиров и растворителей (поиск, сбор, систематизация, обработка и анализ первичных экспериментальных данных) для включения в модель, что подтверждает известный философский афоризм британского статистика Дж. Бокса: «Все модели неверны, но некоторые из них полезны» (1978).

С термодинамической точки зрения, сравнение устойчивости коронатов катионов в разных растворителях требует детального рассмотрения энтальпийных (связевых) и энтропийных (стохастических) вкладов в изменение свободной энергии Гиббса комплексообразования, которые в свою очередь зависят от термодинамических характеристик сольватации (пересольватации) катионов, лигандов и коронатов катионов. Эти вопросы рассмотрены нами в работах, посвященных сольватационно-термодинамическому подходу [65] к исследованию влияния растворителя на силу слабых электролитов и устойчивость катион-краун-эфирных комплексов.

В данной работе демонстрируется применение модельного подхода к установлению статистически значимой связи между физико-химическими свойствами растворителя, катиона, лиганда и устойчивостью комплексов краун-эфиров с катионами на основе совместного использования разведочных регрессионных и нейросетевых алгоритмов для построения прогнозных моделей устойчивости коронатов в разных растворителях.

Компьютерное моделирование позволит, во-первых, прогнозировать устойчивость катионов с молекулами, содержащими краун-эфирные фрагменты, в средах разной природы, а во-вторых, оптимизировать планирование экспериментов в растворителях, в которых комплексообразование краун-эфиров с катионами еще не изучено, либо исследовано недостаточно полно, в частности, в органических и смешанных растворителях.

КОНФЛИКТ ИНТЕРЕСОВ

Автор заявляет об отсутствии конфликта интересов.

СПИСОК ЛИТЕРАТУРЫ

1. *Тьюки Д.* Анализ результатов наблюдений. Разведочный анализ. М.: Мир, 1981. 696 с.
2. *Donoho D.* // J. Comput. Graph. Stat. 2017. Vol. 26. N 4. P. 745. doi 10.1080/10618600.2017.1384734
3. *Брюс П., Брюс Э.* Практическая статистика для специалистов Data Science. Пер. с англ. СПб: БХВ-Петербург, 2018. 304 с.
4. *Chambers J.M.* // Stat. Comput. 1993. Vol. 3. N 4. P. 182. doi 10.1007/bf00141776
5. *Breiman L.* // Stat. Sci. 2001. Vol. 16. N 3. P. 199. doi 10.1214/ss/1009213726
6. *Hill T., Lewicki P.* Statistics: methods and applications: a comprehensive reference for science, industry, and data mining. Tulsa, Okla.: StatSoft. 2006. 832 p.
7. *Dhar V.* // Commun. ACM. 2013. Vol . 56. N 12. P. 64. doi 10.1145/2500499
8. *Guo J., Chen Q., Wang C., Qiu H., Liu B., Jiang Z.-H., Zhang W.* // Anal. Bioanal. Chem. 2015. Vol. 407. N 5. P. 1389. doi 10.1007/s00216-014-8371-x
9. *Komorowski M., Marshall D.C., Salciccioli J.D., Crutain Y.* // Cham: Springer, 2016. Ch. 15. P. 185. doi 10.1007/978-3-319-43742-2_15
10. *Cutcher-Gershenfeld J., Baker K.S., Berente N., Flint C., Gershenfeld G., Grant B., Haberman M., King J.L., Kickpatrick C., Lawrence B., Lewis W., Lenhardt W.C., Mayernik M., McElroy C., Mittleman B., Shin N., Stall S., Winter S., Zaslavsky I.* // Nature. 2017. Vol. 543. P. 615. doi 10.1038/543615a
11. *Ma X., Hummer D., Golden J., Fox P., Hazen R., Morrison S., Downs R.T., Madhikarmi B.L., Wang C., Meyer M.* // ISPRS Int. J. Geo-Inf. 2017. Vol. 6. N 11. P. 368. doi 10.3390/ijgi6110368
12. *Biancolillo A., Marini F.* // Front. Chem. 2018. Vol. 6.

P. 576. doi 10.3389/fchem.2018.00576
13. *Bevilacqua M., Bucci R., Magrì A.D., Magrì, A.L., Nescatelli R., Marini F.* // Chemom. Food Chem. 2013. Vol. 28. P. 171. doi 10.1016/b978-0-444-59528-7.00005-3
14. *Brereton R.G., Jansen J., Lopes J., Marini F., Pomerantsev A., Rodionova O., Roger J.M., Walczak B., Tauler R.* // Anal. Bioanal. Chem. 2018. doi 10.1007/s00216-018-1283-4
15. *Tauler R, Parastar H.* // Angew. Chem. Int. Ed. Engl. 2018. doi 10.1002/anie.201801134
16. *García F.P., García M.A.F., Drożdżak J., Ruiz-Samblás C.* // Environ. Sci. Pollut. Res. 2012. Vol. 19. N 8. P. 3317. doi 10.1007/s11356-012-0849-5
17. *De Klerck K., Vander Heyden Y., Mangelings D.* // J. Chromatogr (A). 2014. Vol. 1326. P. 110. doi 10.1016/j.chroma.2013.12.052
18. *Liu Y., Zhao T., Ju W., Shi S.* // J. Materiomics. 2017. Vol. 3. N 3. P. 159. doi 10.1016/j.jmat.2017.08.002
19. *Wei J.N., Duvenaud D., Aspuru-Guzik A.* // ACS Cent. Sci. 2016. Vol. 2. N 10. P. 725. doi 10.1021/acscentsci.6b00219
20. *Blount D., Banda P., Teuscher C., Stefanovic D.* // Artif. Life. 2017. Vol. 23. N 3. P. 295. doi 10.1162/artl_a_00233
21. *Coley C.W., Jin W., Rogers L., Jamison T.F., Jaakkola T.S., Green W.H., Barzilay R., Jensen K.F.* // Chem. Sci. 2019. Vol. 10. P. 370. doi 10.1039/c8sc04228d
22. *Bonini Neto A., Bonini C.S.B., Reis A.R., Piazentin J.C., Coletta L.F.S., Putti F.F., Heinrichsb R., Moreira A.* // Commun. Soil Sci. Plant Anal. 2019. Vol. 50. N 14. P. 1785. doi 10.1080/00103624.2019.1635144
23. *Meyer J.G., Liu S., Miller I.J., Coon J.J., Gitter A.* // J. Chem. Inf. Model. 2019. Vol. 59. N 10. P. 4438. doi 10.1021/acs.jcim.9b00236
24. *Li H., Zhang Z., Liu Z.* // Catalysts. 2017. Vol. 7. N 10. P. 306. doi 10.3390/catal7100306
25. *Schütt K.T., Arbabzadah F., Chmiela S., Müller K.R., Tkatchenko A.* // Nat. Commun. 2017. Vol. 8. N 13890. P. 1. doi 10.1038/ncomms13890
26. *Molina J., Laroche A., Richard J.-V., Schuller A.-S., Rolando C.* // Front. Chem. 2019. Vol. 7. P. 375. doi 10.3389/fchem.2019.00375
27. *Chen X., Sztandera L., Cartwright H.M.* // Int. J. Intell. Syst. 2007. Vol. 23. N 1. P. 22. doi 10.1002/int.20256
28. *Ye W., Chen C., Wang Z., Chu I.-H., Ong S.P.* // Nat. Commun. 2018. Vol. 9. N 3800. P. 1. doi 10.1038/s41467-018-06322-x
29. *Cova T.F., Canelas_pais A.A.* // Front. Chem. 2019. Vol. 7. P. 809. doi 10.3389/fchem.2019.00809
30. *Alves T.H., Oliveira P., Mota L., Correa C., Abud A.K., Oliveira Junior A.* // Chem. Eng. Trans. 2019. Vol. 74. P. 1483. doi 10.3303/CET1974248
31. *Schmidt J., Marques M.R.G., Botti S., Marques M.A.L.* // npj Comput. Mater. 2019. Vol. 5. N 83. P. 1. doi 10.1038/s41524-019-0221-0
32. *Länge M.* // Soft Comput. 2020. doi 10.1007/s00500-019-04663-3
33. *Bondarev N.V.* // Klin. inform. telemed. 2019. Vol. 14. N 15. P. 141. doi 10.31071/kit2019.15.13
34. *Бондарев Н.В.* // ЖОХ. 2016. Т. 86. № 6. С. 887; *Bondarev N.V.* // Russ. J. Gen. Chem. 2016. Vol. 86. N 6. P. 1221. doi 10.1134/S1070363216060025
35. *Бондарев Н.В.* // ЖОХ. 2017. Т. 87. № 2. С. 207; *Bondarev N.V.* // Russ. J. Gen. Chem. 2017. Vol. 87. N 2. P. 188. doi 10.1134/S1070363217020062
36. *Бондарев Н.В.* // ЖОХ. 2019. Т. 89. № 2. С. 288. doi 10.1134/S0044460X19020197; *Bondarev N.V.* // Russ. J. Gen. Chem. 2019. Vol. 89. N 2. P. 281. doi 10.1134/S1070363219020191
37. *Бондарев Н.В.* // ЖОХ. 2019. Т. 89. № 7. С. 1085. doi 10.1134/S0044460X1907014X; *Bondarev N.V.* // Russ. J. Gen. Chem. 2019. Vol. 89. N 7. P. 1438. doi 10.1134/S1070363219070144
38. *Зенкин А.А.* Когнитивная компьютерная графика. М.: Наука, 1991. 192 с.
39. *Brown F.K.* // Annual Reports in Medicinal Chemistry. 1998. Vol. 33. P. 375. doi 10.1016/s0065-7743(08)61100-8
40. *Leach A.R., Gillet V.J.* An Introduction to Chemoinformatics. Dordrecht: Springer, 2007. 256 p.
41. *Bunin B.A., Siesel A., Morales G.A., Bajorath J.* Chemoinformatics: Theory, Practice, & Products. Dordrecht: Springer, 2007. 295 p.
42. *Baskin V., Varnek V*. Chemoinformatics Approaches to Virtual Screening. Cambridge: RCS Publishing, 2008. 43 p.
43. *Бондарев Н.В.* // ЖОХ. 2020. Т. 90. № 6. С. 953. doi 10.31857/S0044460X20060170; *Bondarev N.V.* // Russ. J. Gen. Chem. 2020. Vol. 90. N 6. P. 1040. doi 10.1134/S1070363220060171
44. *Соловьев И.П.* Дис. ... докт. хим. наук. М., 2007. 350 с.
45. *Хайкин С.* Нейронные сети: полный курс. М.: Издательский дом «Вильямс», 2006. 1104 с.
46. *Гальберштам Н.М., Баскин И.И., Палюлин В.А., Зефиров Н.С.* // Усп. хим. 2003. Т. 72. № 7. С. 706; *Halberstam N.M., Baskin I.I., Palyulin V.A., Zefirov N.S.* // Russ. Chem. Rev. 2003. Vol. 72. N 7. P. 629. doi 10.1070/RC2003v072n07ABEH000754.
47. Халафян А.А. Современные статистические методы медицинских исследований. М.: ЛКИ, 2008. 320 с.
48. *Колмогоров А.Н.* // Докл. АН СССР. 1957. Т. 114. № 5. С. 953.
49. *Ким Дж.-О., Мьюллер Ч.У., Клекка У.Р.* Факторный, дискриминантный и кластерный анализ. М.: Финан-

сы и статистика, 1989. 216 с.
50. *Малхорта Н.К.* Маркетинговые исследования. Практическое руководство. М.: Издательский дом «Вильямс», 2002. 960 с.
51. *Боровиков В.П.* STATISTICA. Искусство анализа данных на компьютере: Для профессионалов. СПб: Питер. 2003. 686 с.
52. *Аксенов С.В., Новосельцев В.Б.* Организация и использование нейронных сетей (методы и технологии). Томск: НТЛ, 2006. 128 с.
53. *Барсегян А.А., Куприянов М.С., Степаненко В.В.* Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. СПб: БХВ-Петербург, 2007. 384 с.
54. *Наследов А.* IBM SPSS Statistics 20 и AMOS: профессиональный статистический анализ данных. СПб: Питер, 2013. 416 с.
55. *Боровиков В.П.* Нейронные сети. Statistica Neural Networks. Методология и технологии современного анализа данных. М.: Горячая линия – Телеком, 2008. 392 с.
56. *Бондарев С.Н., Бондарев Н.В.* // Вест. Харьк. нац. унив. 2010. № 932. Вып. 19(42). С. 70.
57. Бондарев С.Н. *, Зайцева И.С., Бондарев Н.В.* // Бутлеровск. сообщ. 2011. Т. 27. № 14. С. 1.
58. *Бондарев С.Н. , Зайцева И.С., Бондарев Н.В.* // Бутлеровск. сообщ. 2011. Т. 27. № 13. С. 36.
59. *Бондарев С.Н. , Зайцева И.С. , Бондарев Н.В.* // Бутлеровск. сообщ. 2011. Т. 27. № 16. С. 15.
60. *Бондарев Н.В.* // Укр. хим. ж. 1995. Т.61. № 11. С. 14.
61. *Бондарев Н.В.* // Укр. хим. ж. 1998. Т. 64. № 8. С. 85.
62. *Бондарев Н.В.* // ЖОХ. 1999. Т. 69. Вып. 2 С. 229.
63. *Бондарев Н.В.* // ЖФХ. 1999. Т.73. № 6. С. 1019.
64. *Бондарев Н.В.* // ЖОХ. 2006. Т.76. № 1. С. 13; *Bondarev N.V.* // Russ. J. Gen. Chem. 2006. Vol. 76. N 7. P. 11. doi 10.1134/s1070363206010038
65. *Бондарев Н.В.* Термодинамика равновесий. Эффекты среды и нейросетевой анализ. Saarbrucken: LAP LAMBERT Academic Publishing, 2012. 380 с.
66. *Christy F.A., Shrivastav P.S.* // Crit. Rev. Anal. Chem. 2011. Vol. 41. N 3. P. 236. doi 10.1080/10408347.2011.589284
67. *Rodgers M.T., Armentrout P.B.* // Chem. Rev. 2016. Vol. 116. N 9. P. 5642. doi 10.1021/acs.chemrev.5b00688
68. *Marcus Y.* The Properties of Solvents. Chichester: John Wiley & Sons, 1999. Vol. 4. 399 p.
69. *Shannon R.D., Prewitt C.T.* // Acta Crystallogr. (B). 1969. Vol. 25. N 5. P. 925. doi 10.1107/s0567740869003220
70. *Ouchi M., Inoue Y., Kanzaki T., Hakushi T.* // J. Org. Chem. 1984. Vol. 49. N 8. P. 1408. doi 10.1021/jo00182a017
71. *Takeda Y., Mochizuki Y., Tanaka M., Kudo Y., Katsuta S., Ouchi M.* // J. Incl. Phenom. Macrocycl. Chem. 1999. Vol. 33. N 2. P. 217. doi 10.1023/a:1008099827420
72. *Елисеева И.И., Юзбашев М.М.* Общая теория статистики. М.: Финансы и статистика, 2004. 656 с.
73. *Касюк С.Т.* Первичный, кластерный, регрессионный и дискриминантный анализ данных спортивной медицины на компьютере. Челябинск: Уральская Академия, 2015. 160 с.
74. *Лемешко Б.Ю.* Критерии проверки отклонения распределения от нормального закона. Руководство по применению. 2014. Новосибирск: НГТУ, 192 с.
75. *Tong C.* // Am. Stat. 2019. Vol. 73. N s1. P. 246. doi 10.1080/00031305.2018.1518264
76. *Breiman L., Friedman J., Olshen R., Stone C.* Classification and Regression Trees. Belmont: Wadsworth International Group,1984. 358 c.
77. *Nocedal J., Wright S.J.* Numerical Optimization. Dordrecht: Springer, 2006. 683 p.
78. *Al-Baali M., Spedicato E., Maggioni F.* // Optimization Methods and Software. 2013. Vol. 29. N 5. P. 937. doi 10.1080/10556788.2013.856909
79. *Izatt R.M., Bradshaw J.S., Nielsen S.A., Lamb J.D., Christensen J.J., Sen D.* // Chem. Rev. 1985. Vol. 85. N 4. P. 271. doi 10.1021/cr00068a003

Exploration, Regression and Neural Network Analysis of the Stability of Cation Coronates in Some Pure Solvents

N. V. Bondarev*

V.N. Karazin Kharkiv National University, Kharkiv, 61022 Ukraine
** e-mail: bondarev_n@rambler.ru*

Received May 13, 2020; revised July 29, 2020; accepted August 9, 2020

Exploratory, regression, and neural network analysis of the stability constants of crown ether 1:1 complexes [12C4, 16C5, (CH_3)216C5, DB21C7, DB24C8, DCH24C8, DB30C10] with alkaline cations (Li^+, Na^+, K^+, Cs^+, Rb^+), alkaline earth (Ca^{2+}, Sr^{2+}, Ba^{2+}), heavy (Ag^+, Tl^+, Co^{2+}, Cu^{2+}, Pb^{2+}) metals and NH_4^+ in water and organic solvents (methanol, acetonitrile, acetone, *N*,*N*-dimethylformamide, nitrobenzene, nitromethane, 1,2-dichloroethane, propylene carbonate) at 298.15 K obtained by conductometric method was made. Factorial, cluster, discriminant, canonical, decision tree, regression and neural network models of clustering, approximation and prediction of thermodynamic constants of complexation of crown ethers with cations depending on the properties of the ligand, the cation interacting with it, and the solvent used were developed. The trained MLP 7-5-5 Multilayer Perceptron Cluster was 100 percent validated for the *k*-means exploration clustering. Independent data on the stability constants of coronates demonstrate the predictive capabilities of the trained perceptron-approximator MLP 7-7-1.

Keywords: crown ethers, complexation constant, exploratory analysis, multiple linear regression, neural networks, modeling, forecasting