УДК 544.354.081.7:004.021

КЛАССИФИКАЦИЯ И ПРОГНОЗИРОВАНИЕ УСТОЙЧИВОСТИ КОРОНАТОВ НАТРИЯ И КАЛИЯ В ВОДНО-ОРГАНИЧЕСКИХ РАСТВОРИТЕЛЯХ МЕТОДАМИ РАЗВЕДОЧНОГО АНАЛИЗА

© 2019 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 19 августа 2018 г.
После доработки 19 августа 2018 г.
Принято к печати 30 августа 2018 г.

На основе многомерного разведочного анализа данных построены линейные классификационные функции Фишера, канонические линейные дискриминантные функции, деревья (правила) классификации и прогнозирования устойчивости коронатов натрия (18-краун-6Na^+) и калия (18-краун-6K^+) по свойствам водно-органических растворителей: вода–метанол, вода–пропанол-2 и вода–ацетонитрил. Предложенный подход к прогнозированию класса устойчивости коронатов апробирован на независимых экспериментальных данных о константах устойчивости коронатов натрия и калия в смесях вода–диоксан и вода–ацетон. Показано, что построенные классификационные функции и правила обладают довольно высоким прогностическим потенциалом.

Ключевые слова: разведочный анализ, константа комплексообразования, коронаты натрия и калия, водно-органические растворители, эмпирические параметры

DOI: 10.1134/S0044460X19020197

Разведочные методы анализа данных (факторный анализ, кластерный анализ, анализ дискриминантных функций, канонические корреляции, деревья классификации и др.) [1–3] содержат универсальные алгоритмы, предназначенные для выявления закономерностей в многомерных данных, что позволяет исследователю обоснованно выбирать статистическую модель, которая наилучшим образом характеризует реальное поведение исследуемой системы. Поэтому эти методы успешно применяются для решения актуальных задач химии [4–7], медицины [8–11], биологии [12, 13], здравоохранения [14–16], гидрохимии и гидроэкологии [17–19], техногенной безопасности [20] психологии и образования [21].

Цель данной работы заключается в классификации и прогнозировании класса устойчивости коронатов натрия и калия по свойствам водно-органических растворителей алгоритмами методов разведочного анализа. Для анализа использованы экспериментальные данные о константах устойчивости комплексов 18С6 эфира (L) с катионами щелочных металлов Na^+ (NaL) и K^+ (KL), полученные нами ранее при 25°С, в водно-метанольных [22–27], водно-изопропанольных [28, 29] и водно-ацетонитрильных растворителях [30, 31] с шагом 0.1 мол. доли (объем выборки n = 33) и свойства смешанных водно-органических растворителей – ε, E_T, B_{KT} и δ^2, числовые значения которых взяты из источников [32–35].

В комплексообразовании участвуют электронодонор (краун-эфир 18С6) и электроноакцепторы (катионы щелочных металлов), поэтому устойчивость образующегося коронатa в водно-органическом растворителе зависит как от электростатических эффектов (диэлектрическая проницаемость среды, ε), так и от электроноакцепторной и электронодонорной способностей молекул смешанного растворителя. В качестве количественной меры этих свойств растворителя используются сольватохромный параметр Димрота–Райхардта (E_T), отражающий специфическую льюисову кислотность растворителя, и параметр Камлета–Тафта (B_{KT}), отвечающий за его льюисову

Таблица 1. Расчетные и табличные значения критериев проверки гипотезы нормальности распределения переменных[a]

Переменная	Расчетное значение		
	критерий Шапиро–Уилка, $W_{расч}$	критерий Хи-квадрат Пирсона, $\chi^2_{расч}$	критерий Колмогорова–Смирнова, $D_{расч}$
$1/\varepsilon$	0.863	5.352	0.177
E_T	0.976	2.282	0.124
B_{KT}	0.961	1.413	0.126
δ^2	0.957	1.407	0.079
$\lg K_{18C6Na}$	0.965	0.981	0.099
$\lg K_{18C6K}$	0.956	0.318	0.138
Табличное значение (n = 33, p = 0.05)			
	$W_{табл}$	$\chi^2_{табл}$	$D_{табл}$
	0.931	5.991	0.231

[a] n – объем выборки, p – уровень значимости. Если табличное значение $W_{табл}$ меньше расчетного значения $W_{расч}$, то нулевая гипотеза о нормальном распределении не отклоняется при уровне значимости p = 0.05. Если $D_{табл} > D_{расч}$ и $\chi^2_{табл} > \chi^2_{расч}$, то нулевая гипотеза о нормальном распределении переменных принимается при уровне значимости p = 0.05.

основность. Кроме этих параметров, важное значение имеет плотность энергии когезии (когезионное давление, δ^2). Поскольку величина δ^2 фактически является мерой энергии, которую надо затратить на преодоление сил притяжения только между молекулами растворителя, то в общем случае когезионное давление определяет энергию необходимую для образования в растворителе полости, которая может быть заполнена исходными реагентами и продуктами комплексообразования.

Для сопоставительного анализа устойчивости коронатов в разных растворителях термодинамические константы комплексообразования стандартизованы по аквамоляльной концентрационной шкале [37, 38]. Обработку экспериментальных данных проводили в статистических средах STATISTICA 12 и SPSS 23 для Windows [39]. Она включала: 1) первичный анализ данных, вычисление описательных статистик, проверку 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 правил классификации устойчивости коронатов натрия и калия; 7) подтверждение прогностической мощности построенных классификационных функций и правил классификации.

Первичный анализ данных. Проведена проверка характера распределения переменных на нормальность (табл. 1) в соответствии с требованиями ГОСТа [40] по статистикам критериев Шапиро–Уилка, Хи-квадрат Пирсона и Колмогорова–Смирнова. Из анализа данных табл. 1 следует, что эмпирическое распределение анализируемых переменных практически не отличается от нормального.

Факторный анализ. В среде SPSS 23 для обоснования правомерности проведения факторного анализа данных [3] рассчитаны критерии сферичности Бартлетта и адекватности выборки Кайзера–Майера–Олкина (КМО): для $18C6Na^+$ χ^2 = 206.22 (число степеней свободы 10, p = 0.00), КМО = 0.684; для $18C6K^+$ χ^2 = 162.03 (число степеней свободы 10, p = 0.00), КМО = 0.600. Высокие значения критерия Бартлетта ($\chi^2_{табл}$ = 18.307) и КМО (от 0.5 до 1.0) указывают на целесообразность проведения факторного анализа взаимосвязи констант устойчивости коронатов и свойств водно-органических растворителей.

Методом главных компонент по выборочной совокупности шести переменных вычислены корреляционная матрица системы используемых

Таблица 2. Корреляционная матрица переменных

Переменные	Коэффициенты корреляции, n = 33					
	1/ε	E_T	B_{KT}	δ^2	lg$K$	lgK_{NaL}
1/ε	1.00	–0.83	0.90	–0.64	0.35	0.51
E_T	–0.83	1.00	–0.76	0.73	–0.49	–0.65
B_{KT}	0.90	–0.76	1.00	–0.51	0.34	0.41
δ^2	–0.64	0.73	–0.51	1.00	–0.85	–0.97
lgK_{KL}	0.35	–0.49	0.34	–0.85	1.00	0.93
lgK_{NaL}	0.51	–0.65	0.41	–0.97	0.93	1.00

свойств смешанных растворителей и констант устойчивости NaL и KL (табл. 2), ее собственные значения, факторные нагрузки и веса факторов (табл. 3). Проведенный корреляционный анализ показывает, что константы устойчивости коронатов натрия lgK_{NaL} и калия lgK_{KL} проявляют сильную отрицательную взаимосвязь с плотностью энергии когезии δ^2 и умеренную положительную и отрицательную взаимосвязь с 1/ε, B_{KT} и E_T соответственно. Следует отметить, что как отрицательная, так и положительная взаимосвязь между свойствами водно-органических растворителей характеризуется коэффициентами корреляции, превышающими 0.5.

Для отбора латентных факторов (F_1 и F_2) применены метод главных компонент, критерий каменистой осыпи и процедура ортогонального варимакс-вращения факторов [39]. При этом отбрасывали факторы, соответствующие собственные значения которых мало отличались. Величины собственных значений и веса факторов показывают, что значения исследуемых свойств систем константа комплексообразования–свойства среды определяются преимущественно двумя факторами: действием фактора F_1 на 49.43% для комплекса $18C6Na^+$ и на 52.42% для $18C6K^+$; действием фактора F_2 на 44.66% ($18C6Na^+$) и на 39.55% ($18C6K^+$). Анализ признаковой структуры фактора F_1 позволяет заключить, что нагрузка этого фактора определяется свойствами среды 1/ε (–0.9103 и –0.9405), E_T (0.7599 и 0.8120), B_{KT} (–0.9487 и –0.9367). Причем фактор F_1 имеет значимую отрицательную связь с 1/ε и B_{KT} и более слабую, но положительную связь с E_T. Такой характер связи свойств среды с первым фактором позволяет заключить, что фактор F_1 отвечает за влияние свойств водно-органического растворителя на процесс образования комплексов между краун-эфиром и катионами натрия и калия.

Фактор F_2 несет в себе 44.66% ($18C6Na^+$) и 39.55% ($18C6K^+$) информации о рассматриваемых системах. Анализ признаковых нагрузок этого фактора показывает, что он имеет значимую отрицательную связь с константами устойчивости коронатов натрия lgK_{NaL} = –0.9691 и калия lgK_{KL} = –0.9641, а также значимые положительные связи с плотностью энергии когезии δ^2 (0.9300 и 0.8728). Такая признаковая структура фактора F_2 дает основание полагать, что связь между устойчивостью коронатов и плотностью энергии когезии рассматриваемых водно-органических растворителей носит антибатный характер.

Таблица 3. Факторные нагрузки, собственные значения и веса факторов

Параметр	$18C6Na^+$		$18C6K^+$	
	факторные нагрузки			
	F_1	F_2	F_1	F_2
1/ε	–0.9103	–0.3346	–0.9405	–0.2477
E_T	0.7599	0.5310	0.8120	0.4438
B_{KT}	–0.9487	–0.1870	–0.9367	–0.1665
δ^2	0.3476	0.9300	0.4278	0.8728
lg$K$	–0.2110	–0.9691	–0.1298	–0.9641
Собственные значения	2.4715	2.2329	2.6212	1.9773
Вес фактора, %	49.4300	44.6600	52.4200	39.5500

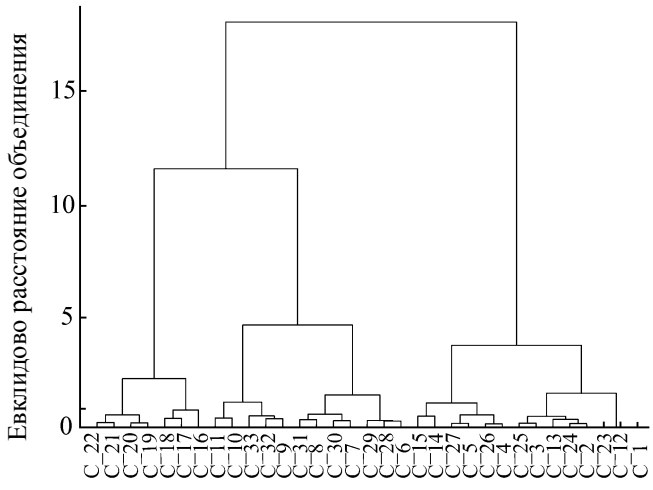

Рис. 1. Дендрограмма иерархической классификации устойчивости коронатa калия.

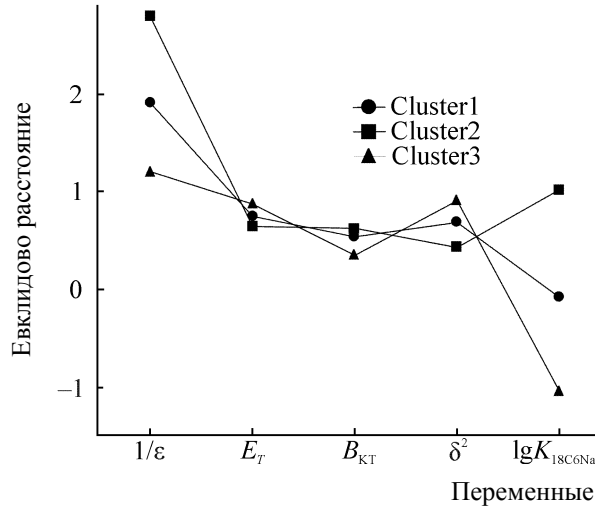

Рис. 2. График средних для трех кластеров устойчивости коронатa натрия.

Кластерный анализ. В работе реализованы два метода кластерного анализа: агломеративный – объединение, или дерево кластеризации, и дивизивный – кластеризация *k*-средними.

На рис. 1 приведена дендрограмма иерархической классификации устойчивости коронатa калия (lg*K*: C_1–C_33) по свойствам водно-органических растворителей. Кластеризация выполнена методом Варда с использованием евклидового расстояния в качестве метрики пространства. На расстоянии, равном 5, существуют 3 кластера; при увеличении расстояния до 13 количество кластеров становится равным двум, а на расстоянии, равном 21, остался один кластер.

На рис. 2 приведен график средних значений по каждому кластеру, который показывает наибольшее различие между тремя выделенными кластерами. Точки их средних значений по пяти переменным находятся на самых больших расстояниях друг от друга, что особенно характерно для параметров 1/ε и lg*K*. Результаты дисперсионного анализа свидетельствуют (табл. 4), что разделение на кластеры проведено успешно. Уровень значимости *p* у критерия Фишера значительно меньше 0.05 для всех переменных и наблюдаемый критерий Фишера больше критического $F_{набл} > F_{кр}$.

Сравнивая результаты кластеризации по алгоритмам *k*-средних и древовидной кластеризации (табл. 5 на примере коронатa калия) можно заключить, что содержимое первых кластеров совпадает, а в содержимом вторых и третьих кластеров есть отличия, отмеченные полужирным шрифтом.

Выполненный кластерный анализ данных позволяет интерпретировать содержимое трех кластеров (классов): кластер 1 – умеренно устойчивые коронаты в смешанных растворителях промежуточного состава ($\lg K_{NaL}$ = 1.5–2.5, $\lg K_{KL}$ = 3.1–3.9); кластер 2 – устойчивые комплексы в растворителях с большим содержанием органического компонента и в чистых неводных растворителях ($\lg K_{NaL}$ = 2.6–4.3, $\lg K_{KL}$ = 4.0–5.2); кластер 3 – слабо устойчивые коронаты в воде и смешанных растворителях с большим содержанием воды ($\lg K_{NaL}$ = 0.5–1.4, $\lg K_{KL}$ = 2.0–3.0).

Любая кластеризация всегда носит субъективный характер, потому что выполняется на основе конечного набора переменных и разными алгоритмами, каждый из которых имеет свои достоинства, недостатки и ограничения [1]. Поэтому для подтверждения результатов кластерного анализа в работе проведен дискриминантный и канонический анализ данных, а также построены деревья решений (правила классификации).

Дискриминантный анализ. Цель дискриминантного анализа состояла в том, чтобы на основе независимых параметров (свойств водно-органических растворителей) классифицировать константы устойчивости коронатов натрия или калия, то есть отнести их к одному из трех классов, выделенных итерационным методом *k*-средних (табл. 6).

Таблица 4. Результаты дисперсионного анализа переменных комплексообразования 18С6 эфира с катионами натрия и калия методом *k*-средних[a]

Переменная	σ_1^2	ν_1	σ_2^2	ν_2	*F*	*p*
	18C6Na$^+$					
1/ε	14.60	2	11.78	30	18.60	0.000006
E_T	0.31	2	0.26	30	17.71	0.000008
B_{KT}	0.46	2	0.72	30	9.59	0.000601
δ^2	1.33	2	0.31	30	64.48	0.000000
lgK_{NaL}	24.57	2	7.43	30	49.64	0.000000
	18C6K$^+$					
1/ε	22.85	2	3.54	30	96.94	0.000000
E_T	0.31	2	0.26	30	17.95	0.000007
B_{KT}	0.77	2	0.40	30	28.80	0.000000
δ^2	1.08	2	0.55	30	29.40	0.000000
lgK_{KL}	23.88	2	8.12	30	44.14	0.000000

[a] σ_1^2 – межгрупповая дисперсия; σ_2^2 – внутригрупповая дисперсия; ν_1, ν_2 – степени свободы; $F(2, 30)$ – наблюдаемый критерий Фишера [$F_{кр}(2, 30, p\ 0.05) = 2.04$]; *p* – наблюдаемый уровень значимости.

Таблица 5. Результаты кластерного анализа

Разбиение выборки (33 наблюдения) на 3 кластера		
Кластер 1	Кластер 2	Кластер 3
Агломеративная кластеризация (7 + 12 + 14)		
16, 17, 18, 19, 20, 21, 22	**6**, **28**, 29, 7, 30, 8, 31, 9, 32, 33, 10, 11	1, 12, 23, 2, 24, 13, 3, 25, 4, 26, 5, 27, 14, 15
Кластеризация *k*-средними (7 + 10 + 16)		
16, 17, 18, 19, 20, 21, 22	7, 8, 9, 10, 11, 29, 30, 31, 32, 33	1, 2, 3, 4, 5, **6**, 12, 13, 14, 15, 23, 24, 25, 26, 27, **28**

Таблица 6. Результаты дискриминантного анализа (алгоритм – переменные в модели)

Параметр	Группирующая переменная: Кластер NaL (3 кластера); Λ-Уилкса: 0.120; $F_{набл}(6, 56) = 17.620$, $p < 0.000$; $F_{кр}(6, 56) = 2.25$			
	Λ-Уилкса	частная Λ-Уилкса	$F_{иск}$	*p*-уровень
δ^2	0.363	0.330	28.415	0.000
1/ε	0.153	0.784	3.858	0.033
B_{KT}	0.142	0.847	2.538	0.097
	Группирующая переменная: Кластер KL (3 кластера); Λ-Уилкса: 0.024; $F_{набл}(6, 56) = 50.616$, $p < 0.000$; $F_{кр}(6, 56) = 2.25$			
1/ε	0.133	0.183	62.554	0.000
δ^2	0.128	0.189	59.984	0.000
B_{KT}	0.028	0.870	2.091	0.142

Таблица 7. Матрица классификации для моделей Кластер NaL и Кластер KL[a]

Кластер	% правильной классификации	Cluster1NaL	Cluster2NaL	Cluster3NaL
Кластер NaL				
Cluster1NaL	100	10	0	0
Cluster2NaL	100	0	12	0
Cluster3NaL	90.91	1	0	10
Всего, %	96.97	11	12	10
Кластер KL				
Cluster1NaL	100	7	0	0
Cluster2KL	100	0	10	0
Cluster3KL	93.75	0	1	15
Всего, %	96.97	7	11	15

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

Для дискриминации констант устойчивости коронатов использован линейный дискриминантный анализ Фишера, реализованный в статистическом пакете STATISTICA 12. Представляют интерес основные сведения о методе анализа, а также о переменных, включенных в дискриминантную модель, и значения статистических показателей (табл. 6). Значение Λ-Уилкса для модели Кластер NaL равно 0.120, а для модели Кластер KL – 0.024. Таким образом, обе модели демонстрируют хорошую дискриминацию констант устойчивости по трем кластерам, но дискриминирующая мощность второй модели в 5 раз выше, чем у первой.

Значения $F_{набл}$-статистики [$F_{набл}(6, 56) = 17.620$, $p < 0.000$ для первой модели и $F_{набл}(6, 56) = 50.616$, $p < 0.000$ для второй модели], связанной с величиной Λ-Уилкса, свидетельствуют о статистической значимости моделей дискриминации, так как $F_{набл}(6, 56) > F_{кр}(6, 56)$.

Из данных табл. 6 следует, что только две переменные ($1/\varepsilon$ и δ^2) наиболее информативны: чем больше значение Λ-Уилкса, тем более желательна эта переменная в процедуре дискриминации. Однако, если в первой модели Кластер NaL эти переменные желательны в процедуре дискриминации в следующем порядке: δ^2, $1/\varepsilon$, то во второй модели Кластер KL переменные по силе дискриминации расположены в обратном порядке. Частная Λ-Уилкса, характеризующая единичный вклад соответствующей переменной в разделительную силу модели, подтверждает этот вывод. Чем меньше значение частной Λ-Уилкса, тем больший вклад этой переменной в общую дискриминацию. Наряду с этим, чем меньше значение критерия Фишера $F_{искл}$ и больше p-уровень (табл. 6), тем менее желательны переменные в модели дискриминации. Поэтому переменная $B_{КТ}$ в обеих дискриминационных моделях менее информативна, так как для нее $p > 0.05$.

Числовые значения линейных классификационных функций можно рассчитать по формулам (1)–(3) для модели Кластер NaL и (4)–(6) для модели Кластер KL.

$$\text{Cluster1NaL} = -27.947 + 63.800\delta^2 + 0.258(1/\varepsilon) + 16.875B_{КТ}, \quad (1)$$

$$\text{Cluster2NaL} = -18.312 + 37.767\delta^2 + 5.991(1/\varepsilon) + 2.491B_{КТ}, \quad (2)$$

$$\text{Cluster3NaL} = -42.167 + 87.120\delta^2 - 0.144(1/\varepsilon) + 8.568B_{КТ}, \quad (3)$$

$$\text{Cluster1KL} = -293.310 + 143.620(1/\varepsilon) + 286.877\delta^2 - 101.121B_{КТ}, \quad (4)$$

$$\text{Cluster2KL} = -116.617 + 88.495(1/\varepsilon) + 184.117\delta^2 - 58.796B_{КТ}, \quad (5)$$

$$\text{Cluster3KL} = -142.763 + 93.421(1/\varepsilon) + 214.505\delta^2 - 62.089B_{КТ}. \quad (6)$$

Таблица 8. Характеристика извлеченных канонических корней[a]

Корень	Co	R	R^2	Λ	χ^2	ν	p
0	13.52	0.965	0.931	0.024	107.87	6	0.000
1	1.84	0.805	0.648	0.352	30.27	2	0.000

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

Подставив в эти уравнения значения свойств водно-органических растворителей, которые не использовались при построении линейных классификационных функций, можно предсказать класс устойчивости коронатов натрия или калия в этих растворителях по рассчитанному значению линейной классификационной функции. Константа устойчивости при этом будет отнесена к конкретному классу (первому, второму или третьему) по наибольшему числовому значению линейных классификационных функций.

Матрица классификации (табл. 7) позволяет оценить качество линейных классификационных функций. На диагонали матрицы содержится количество констант устойчивости коронатов натрия или калия, корректно классифицированных в кластеры. Как видно из таблицы, общий вклад правильной классификации в двух моделях составляет 96.97%. К кластеру Cluster3NaL правильно отнесены 10 констант устойчивости из 11, что составляет 90.91% правильной классификации. Одна константа устойчивости коронатa натрия ошибочно отнесена к первому классу Cluster1NaL. К кластеру Cluster3KL правильно отнесены 15 констант устойчивости из 16, процент правильной классификации составляет 93.75%. При этом одна константа устойчивости ошибочно классифицирована алгоритмом дискриминантного анализа во второй кластер Cluster2KL. Таким образом, данные табл. 7 свидетельствуют о достаточно высокой дискриминирующей способности моделей и подтверждают результаты дивизивной кластеризации констант устойчивости коронатов методом *k*-средних.

Канонический анализ. Результаты канонического анализа позволили определить (табл. 8) вклад двух канонических линейных дискриминантных функций в дисперсию исследованных независимых параметров – свойств водно-органических растворителей.

Первая каноническая линейная дискриминантная функция извлеченных канонических корней описывает наибольшую часть дисперсии свойств водно-органических растворителей. Вторая каноническая линейная дискриминантная функция описывает наибольшую часть дисперсии свойств, оставшихся не объясненными первой канонической линейной дискриминантной функцией. В табл. 8 эти значения канонической корреляции равны 0.965 и 0.805.

Следовательно, анализ канонической модели, включающей два канонических корня (табл. 8), описывающих структуру зависимости исследуемой совокупности факторов (свойств водно-органических растворителей), свидетельствует о том, что имеющая между ними многомерная взаимосвязь, может быть описана с позиции двух наиболее информативных канонических функций, объясняющих 93.1 и 64.8% всей дисперсии исследуемых переменных.

Уравнения для расчета двух канонических линейных дискриминантных функций для каждой дискриминантной модели имеют следующий вид:

– для модели Кластер NaL

$$D_{1,\mathrm{NaL}} = 4.227 - 8.738\delta^2 + 1.128(1/\varepsilon) - 1.264B_{\mathrm{KT}}, \quad (7)$$

$$D_{2,\mathrm{NaL}} = 0.007 - 1.670\delta^2 - 2.683(1/\varepsilon) + 12.732B_{\mathrm{KT}} \quad (8)$$

– для модели Кластер KL

$$D_{1,\mathrm{KL}} = -16.508 + 6.100(1/\varepsilon) + 10.021\delta^2 - 4.715B_{\mathrm{KT}}, \quad (9)$$

$$D_{2,\mathrm{KL}} = 6.649 - 0.582(1/\varepsilon) - 8.415\delta^2 + 0.278B_{\mathrm{KT}}. \quad (10)$$

Таблица 9. Средние канонических переменных (центроиды кластеров)

Кластер	Кластер NaL		Кластер KL	
	$D_{1,\mathrm{NaL}}$	$D_{2,\mathrm{NaL}}$	$D_{1,\mathrm{KL}}$	$D_{2,\mathrm{KL}}$
Cluster1	–0.308	0.598	6.744	0.160
Cluster2	2.828	–0.221	–2.137	1.796
Cluster3	–2.805	–0.303	–1.615	–1.193

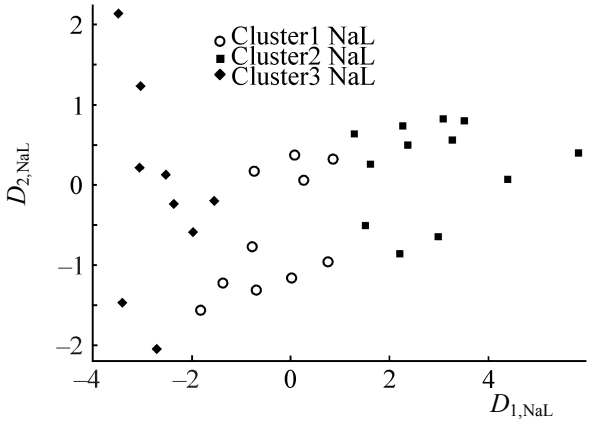

Рис. **3.** Распределение констант устойчивости коронатов натрия по трем кластерам в координатах первой и второй канонических линейных дискриминантных функций.

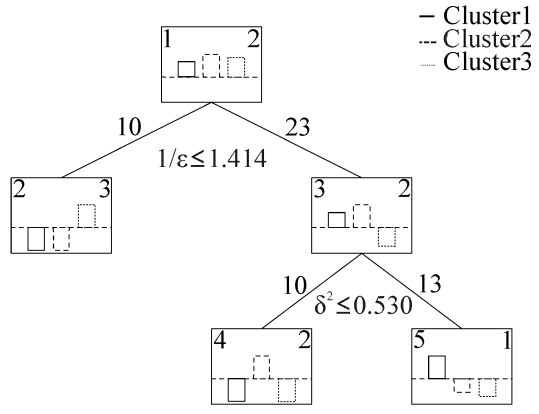

Рис. 4. Граф дерева классификации устойчивости короната натрия.

Функции $D_{1,NaL}$ и $D_{1,KL}$ в каждой из двух моделей наиболее важны, так как ответственны за 97.3 и 88.0% объясненной дисперсии свойств водно-органических растворителей.

Константа устойчивости комплексов, для которой по свойствам водно-органического растворителя рассчитаны значения канонических линейных дискриминантных функций D_1 и D_2, классифицируется в кластер по минимальному расстоянию до соответствующего центроида кластера. Поэтому в табл. 9 приведены координаты центроидов трех кластеров в каждой из рассматриваемых моделей. В каждой модели три кластера достаточно отчетливо дискриминируются между собой первой дискриминантной функцией D_1. В качестве примера приведен рис. 3.

Деревья принятия решений. Построены классификационные модели, позволяющие прогнозировать класс устойчивости коронатов натрия и калия по свойствам водно-органических растворителей (разделять константы устойчивости между тремя кластерами) и решать, какое свойство будет наиболее целесообразным признаком классификации.

Для принятия решения о целесообразности использования того или иного свойства растворителя для классификации устойчивости коронатов натрия или калия выбран алгоритм CART (Classification and Regression Tree). Задачей алгоритма является построение бинарных деревьев решений путем разделения на каждом шаге множества констант устойчивости на две ветви. По одной ветви идут те константы устойчивости, для которых правило выполняется (левый потомок), по другой – те, для которых правило не выполняется (правый потомок).

На рис. 4 приведен граф дерева классификации устойчивости короната натрия. Первоначально все 33 константы алгоритм приписывает к корневой вершине ветвления. На рисунке она помечена цифрой 1 в левом верхнем углу корня дерева. Все 33 константы устойчивости короната натрия предварительно классифицируются как Cluster2 (на это указывает цифра 2 в правом верхнем углу вершины). Cluster2 был выбран алгоритмом для начальной классификации потому, что число констант устойчивости во втором кластере немного большее (12), чем в первом (10) и третьем (11) кластерах (на это указывает гистограмма, изображенная внутри корневой вершины).

Корневая вершина разветвляется на две новые вершины. Текст под корневой вершиной описывает схему ветвления. Если константы устойчивости короната натрия характеризуются свойством растворителя $1/\varepsilon$ меньшим или равным 1.67, то они отнесены алгоритмом к вершине номер 2 и предположительно классифицированы как Cluster3, а константы устойчивости в растворителях с $1/\varepsilon > 1.67$ приписаны к вершине 3 и классифицированы как Cluster2. Числа 10 и 23 над вершинами 2 и 3 соответственно обозначают число констант устойчивости в этих двух дочерних вершинах после первого ветвления родительской корневой вершины. Затем точно также разветвляется

Таблица 10. Структура дерева классификации для модели Кластер KL

Вершина	Левая вершина	Правая вершина	Claster 1	Claster 2	Claster 3	Предсказанный кластер	Константа ветвления	Переменная ветвления
1	2	3	7	10	16	3	1.67	1/ε
2			0	0	15	3		–
3	4	5	7	10	1	2	0.64	B_{KT}
4			0	10	0	2		–
5			7	0	1	1		–

Таблица 11. Рассчитанные значения (линейных классификационных функций по уравнениям (1)–(6) и предсказанные кластеры

Мол. доля S	Cluster1NaL	Кластер	Cluster2NaL	Кластер	Cluster3NaL	Кластер
	Вода–диоксан (S)					
0.0	39.4		25.9		46.5	3
0.1	37.1		27.1		40.6	3
0.2	33.5		30.0		34.0	3
	Вода–ацетон (S)					
0.0	39.4		25.9		46.5	3
0.1	37.7		24.8		41.0	3
0.2	34.1		23.5		34.6	3
0.3	29.3	1	22.3		27.6	
0.4	24.1	1	21.2		20.4	
Мол. доля S	Cluster1 KL	Кластер	Cluster2 KL	Кластер	Cluster3 KL	Кластер
	Вода–диоксан (S)					
0.0	118.0		144.8		153.4	3
0.1	169.6		176.6		185.3	3
0.2	277.4	1	242.8		253.4	
0.3	426.2	1	334.0		348.0	
0.4	612.7	1	448.3		467.0	
	Вода–ацетон (S)					
0.0	117.8		144.7		153.3	3
0.1	107.4		138.4		144.9	3
0.2	114.0		142.2		147.3	3
0.3	131.9		152.7		156.7	3
0.4	154.9		166.3		169.4	3
0.5	177.2		179.5		181.7	3
0.55	186.3		184.9		186.5	3

вершина 3. В результате 10 констант устойчивости в растворителях со значением плотности энергии когезии δ^2 меньшими или равными 0.530 приписываются алгоритмом к вершине 4 и классифицируются как Cluster2, а остальные константы устойчивости коронатa натрия в растворителях с $\delta^2 > 0.530$ – к вершине 5 и классифицируются как Cluster1. Таким образом, точность классификации алгоритмом CART составляет 90.9%. Три константы устойчивости классифицированы ошибочно – две константы устойчивости из второго кластера и одна константа устойчивости из третьего кластера отнесены к первому кластеру.

В табл. 10 приведена структура дерева классификации устойчивости коронатa калия по свойствам водно-органических растворителей. В этой таблице результатов кластеризации вершины 2, 4 и 5 отмечены как терминальные (–), так как в них не происходит ветвление. Дерево классификации на 97.0% подтвердило результаты кластерного анализа устойчивости коронатов калия, проведенного итерационным методом *k*-средних. Только одна константа третьего класса ошибочно отнесена алгоритмом CART в первый кластер.

По аналогии с ранее рассмотренной моделью Кластер NaL для модели Кластер KL также можно построить правило классификации (табл. 10). Если значение $1/\varepsilon$ растворителей меньше или равно 1.67, константы устойчивости коронатa калия в этих растворителях классифицируют как Cluster3. Если значение $1/\varepsilon > 1.67$, а значение параметра Камлета–Тафта растворителей меньше или равно 0.64, константы устойчивости коронатa калия в таких растворителях классифицируют как Cluster2, а в растворителях с $B_{KT} > 0.64$ константы классифицируют как Cluster1.

Предсказательный потенциал алгоритмов разведочных методов анализа. Критерием достоверности построенных классификационных функций и правил является проверка их прогностических возможностей. Проверка предсказательной способности использованных в работе алгоритмов разведочного анализа была реализована прогнозированием класса устойчивости коронатов натрия и калия по свойствам водно-диоксановых [41] и водно-ацетоновых [42–44] растворителей с использованием полученных линейных классификационных функций, числовые значения которых рассчитаны по формулам (1)–(6) и приведены в табл. 11, канонических линейных дискриминантных функций, числовые значения которых рассчитаны по формулам (7)–(10) и приведены в табл. 12, а также построенных правил классификации устойчивости коронатов (табл. 13). Следует отметить, что в литературе есть данные о константах устойчивости коронатов натрия и калия в смешанных растворителях вода–диоксан и вода–ацетон с содержанием органического компонента только до 0.55 мол. доли [41–44].

Таблица 12. Рассчитанные по уравнениям (7)–(10) значения канонических линейных дискриминантных функций и предсказанные кластеры

Мол. доля S	$D_{1,NaL}$	$D_{2,NaL}$	Кластер
Вода–диоксан (S)			
0.0	–3.6	–1.9	3
0.1	–2.4	–1.4	3
0.2	–0.7	–2.7	3
Вода–ацетон (S)			
0.0	–3.6	–1.9	3
0.1	–2.9	0.2	3
0.2	–2.0	0.8	3
0.3	–1.0	0.5	1
0.4	0.1	–0.3	1
0.5	1.2	–0.9	1
Мол. доля S	$D_{1,KL}$	$D_{2,KL}$	Кластер
Вода–диоксан (S)			
0.0	–1.3	–2.3	3
0.1	1.0	–1.9	3
Вода–ацетон (S)			
0.0	–1.3	–2.3	3
0.1	–1.6	–1.7	3
0.2	–1.2	–1.1	3
0.3	–0.3	–0.6	3
0.4	0.9	–0.1	3
0.5	2.0	0.4	3
0.55	2.4	0.7	3

Из анализа данных, приведенных в табл. 11–13, следуют важные выводы. Во-первых, свойства водно-органических растворителей, выбранные для выявления особенностей влияния среды на

Таблица 13. Предсказательная способность деревьев решений при классификации устойчивости коронатов натрия и калия в смесях вода–диоксан и вода–ацетон

Мол. доля S	1/ε	lgK_{KL}	Правило, кластер		1/ε	lgK_{KL}	Правило, кластер	
	вода–диоксан (S)				вода–ацетон (S)			
0	1.000	2.04	1/ε ≤ 1.67	3	1.000	2.04	1/ε ≤ 1.67	3
0.1	1.645	2.67	1/ε ≤ 1.67	3	1.237	2.53	1/ε ≤ 1.67	3
0.2	2.625	3.32			1.514	2.99	1/ε < 1.67	3
Мол. доля S	1/ε	lgK_{NaL}	Правило, кластер		1/ε	lgK_{NaL}	Правило, кластер	
0	1.000	0.52	1/ε ≤ 1.41	3	1.000	0.52	1/ε ≤ 1.41	3
0.1	1.645	1.38			1.237	1.29	1/ε ≤ 1.41	3

устойчивость коронатов натрия и калия, являются статистически значимыми. Во-вторых, построенные линейные классификационные функции Фишера и канонические линейные дискриминантные функции обладают бóльшим прогностическим потенциалом (табл. 11, 12), чем деревья классификации (табл. 13). В-третьих, для повышения прогностической мощности алгоритмов разведочного анализа необходимо пополнение массива данных как по константам устойчивости коронатов, так и по свойствам водно-органических растворителей. В-четвертых, представляется актуальным использование результатов первичного разведочного анализа данных при проведении множественного регрессионного анализа и нейросетевого моделирования [7] для решения задач нелинейной регрессии (аппроксимации) и предсказания класса устойчивости коронатов натрия и калия по свойствам водно-органических растворителях.

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов.

СПИСОК ЛИТЕРАТУРЫ

1. *Тьюки Дж.* Анализ результатов наблюдений. Разведочный анализ. М.: Мир, 1981. 696 с.
2. *Dillon W.R., Goldstein M.* Multivariate Analysis: Methods and Applications. New York: Wiley, 1984. 587 p.
3. *Ким Дж.-О., Мьюллер Ч.У., Клекка У.Р., Олдендерфер М.С., Блэшфилд Р.К.* Факторный, дискриминантный и кластерный анализ. М.: Финансы и статистика, 1989. 215 с.
4. *Jaumot J., Eritja R., Gargallo R.* // Anal. Bioanal. Chem. 2011. Vol. 399. N 6. P.1983. doi 10.1007/s00216-010-4310-7
5. *Fang S.C., Chang I.-C., Yu, T.Y.* // J. Coast. Res. 2015. Vol. 31. N 5. P. 1183. doi 10.2112/JCOASTRES-D-13-00179.1
6. *Бондарев Н.В.* // ЖОХ. 2016. Т. 86. Вып. 6. С. 887; *Bondarev N.V.* // Russ. J. Gen. Chem. 2016. Vol. 86. N 6. P. 1221. doi 10.1134/S1070363216060025
7. *Бондарев Н.В.* // ЖОХ. 2017. Т. 87. Вып. 2. С. 207; *Bondarev N.V.* // Russ. J. Gen. Chem. 2017. Vol. 87. N 2. P. 188. doi 10.1134/S1070363217020062
8. *Hauben M., Hung E., Hsieh W.-Y.* // Ther. Adv. Drug Saf. 2017. Vol. 8. N 1. P. 4. doi 10.1177/2042098616670799
9. *Gorrochategui E., Jaumot J., Lacorte S., Tauler R.* // TrAC Trends Anal. Chem. 2016. Vol. 82. P. 425. doi 10.1016/j.trac.2016.07.004
10. *Zarei K., Taheri F.* // Russ. Chem. Bull. 2016. Vol. 65. N 4. P. 1131. doi 10.1007/s11172-016-1424-x
11. *Ташкинов А.А., Вильдеман А.В., Бронников В.А.* // Рос. ж. биомех. 2008. Т. 12. № 4 (42). С. 84.
12. *Wiles L., Brodahl M.* Weed Science. 2004. Vol. 52. N 6. P. 936. doi 10.1614/WS-03-068R
13. *Kaneene J.B., Miller R.A., Sayah R., Johnson Y.J., Gilliland D., GardinerJ.C.* // Appl. Environ. Microbiol. 2007. Vol. 73. N 9. P. 2878. doi 10.1128/AEM.02376-06
14. *Eisenberg J.N.S., McKone T.E.* // Environ. Sci. Technol. 1998. Vol. 32. N 21. P.3396. doi 10.1021/es970975s
15. *Qiu S., Gao L., Wang J.* // J. Food Eng. 2015. Vol. 144. P. 77. doi 10.1016/j.jfoodeng.2014.07.015
16. *Халафян А.А., Темердашев З.А., Гугучкина Т.И., Якуба Ю.Ф.* // Аналитика и контроль. 2017. Т. 21. № 2. С. 161. doi 10.15826/analitika.2017.21.2.010
17. *Kumar M., Singh Ya.* // J. Water Resource Protect. 2010. Vol. 2. N 10. P. 860. doi 10.4236/jwarp.2010.210102
18. *Moghimi H.* // Open J. Geol. 2017. Vol. 7. N 6 P. 830.

doi 10.4236/ojg.2017.76057

19. *Кошелева Н.Е., Власов Д.В., Корляков И.Д., Касимов Н.С.* // Вестн. Пермск. НИПУ. Прикладная экология. Урбанистика. 2018. № 1. С. 36. doi 10.15593/2409-5125/2018.01.03
20. *Williams B.A., Onsman A., Brown G.T.* // J. Emerg. Prim. Health Care. 2010. Vol. 8. N 3. P. 1.
21. *Тютюник В.В., Бондарев Н.В., Шевченко Р.И., Черногор Л.Ф., Калугин В.Д.* // Геоінформатика. Київ: Інститут геологічних наук НАН України, 2014. № 4(52). С. 63.
22. *Переселко В.Ф., Шевченко И.А., Жолновач А.М., Бондарев Н.В.* // ЖОХ. 1995. Т. 65. Вып. 3. С. 363.
23. *Переселко В.Ф., Липовецкая Е.Е., Кабакова Е.Н., Бондарев Н.В.* // ЖОХ. 1995. Т. 65. Вып. 3. С. 366.
24. *Кабакова Е.Н., Шевченко И.А., Жолновач А.М., Бондарев Н.В.* // ЖОХ. 1996. Т. 66. Вып. 2. С. 208.
25. *Кабакова Е.Н., Переверзев А.Ю., Бондарев Н.В.* // Укр. хим. ж. 1996. Т. 62. № 1. С. 21.
26. *Липовецкая Е.Е., Кабакова Е.Н., Бондарев Н.В.* /// ЖОХ. 1996. Т. 66. Вып. 2. С.204.
27. *Кабакова Е.Н., Бондарев Н.В.* // ЖФХ. 1998. Т. 72. № 7. С. 1196.
28. *Диди Ю., Бондарев Н.В.* // ЖОХ. 1996. Т. 66. Вып. 8. С. 1267.
29. *Диди Ю., Цурко Е.Н., Бондарев Н.В.* // ЖОХ. 1997. Т. 67. Вып. 6. С. 885.
30. *Ельцов С.В., Юрченко В.А., Бондарев Н.В.* // ЖОХ. 1996. Т. 66. Вып. 4. С. 549.
31. *Ельцов С.В., Кабакова Е.Н., Бондарев Н.В.* // Укр. хим. ж. 1998. Т.64. №4. С.84.
32. *Крестов Г.А., Афанасьев В.Н., Агафонов А.В., Гольдштейн И.П., Федотов А.Н., Кукушкин Ю.Н., Кукушкин М.Ю., Шорманов В.А., Березин М.Б., Павлов Н.Н., Артемов А.В., Вайнштейн Э.Ф.* Комплексообразование в неводных растворах. М.: Наука, 1989. 256 с.
33. *Зайцева И.С., Ельцов С.В., Кабакова Е.Н., Бондарев Н.В.* // ЖОХ. 2003. Т. 73. Вып. 7. С. 1079; *Zaitseva I.S., El'tsov S.V., Kabakova E.N., Bondarev N.V.* // Russ. J. Gen. Chem. 2003. Vol. 73. N 7. P. 1021. doi 10.1023/B:RUGC.0000007603.08621.7e
34. *Афанасьев В.Н., Ефремова Л.С., Волкова Т.В.* Физикохимические свойства бинарных растворителей. Водосодержащие системы. Иваново: ИХНР, 1988. 413 с.
35. *Kalidas C., Hefter G., Marcus Y.* // Chem. Rev. 2000. Vol. 100. N 3. P. 819. doi 10.1021/cr980144k
36. *Райхардт К.* Растворители и эффекты среды в органической химии. М.: Мир, 1991. 763 с.
37. *Бондарев Н.В.* Термодинамика равновесий. Эффекты среды и нейросетевой анализ. Saarbrücken: LAP LAMBERT Academic Publishing, 2012. 380 с.
38. *Tsurko E.N., Bondarev N.V.* // J. Mol. Liquids. 2007. N 131–132. P. 151. doi 10.1016/j.molliq.2006.08.051
39. *Боровиков В.* STATISTICA. Искусство анализа данных на компьютере. СПб: Питер, 2003. 686 с.
40. ГОСТ Р ИСО 5479-2002. Статистические методы. Проверка отклонения распределения вероятностей от нормального распределения.
41. *Зайцева И.С., Григорьева Н.Ю., Ельцов С.В., Бондарев Н.В.* // ЖОХ. 2001. Т. 71. Вып. 4. С. 544; *Zaitseva I.S., Grigor'eva N.Yu., El'tsov S.V., Bondarev N.V.* // Russ. J. Gen. Chem. 2001. Vol. 71. N 4. P. 505. doi 10.1023/A:1012310613474
42. *Кабакова Е.Н., Бондарев Н.В.* // ЖНХ. 1997. Т. 42. № 7. С. 1208.
43. *Кабакова Е.Н., Бондарев Н.В.* // ЖНХ. 1998. Т. 43. № 5. С. 820.
44. *Кабакова Е.Н., Цурко Е.Н., Бондарев Н.В.* // Укр. хим. ж. 1998. Т.64. № 9. С.18.

Classification and Prediction of Sodium and Potassium Coronates Stability in Aqueous Organic Solvents by Exploratory Data Analysis Methods

N. V. Bondarev*

V.N. Karazin Kharkiv National University, pl. Svobody 4, Kharkiv, 61022 Ukraine
** e-mail: bondarev_n@rambler.ru*

Received August 19, 2018; revised August 19, 2018; accepted August 30, 2018

Based on the multivariate exploratory data analysis, linear Fisher classification functions, canonical linear discriminant functions, trees (rules) of classification and prediction of the stability of sodium (18-crown-6Na^+) and potassium coronates (18-crown-6K^+) according to the aqueous organic solvents (water–methanol, water–propan-2-ol, water–acetonitrile) properties were constructed. The proposed approach to predicting the stability class of coronates was tested on independent experimental data on the stability constants of sodium and potassium coronates in a water–dioxane and water–acetone mixtures. The constructed classification functions and rules were found to have a rather high predictive potential.

Keywords: exploratory data analysis, complexation constant, sodium and potassium coronates, aqueous organic solvents, empirical parameters