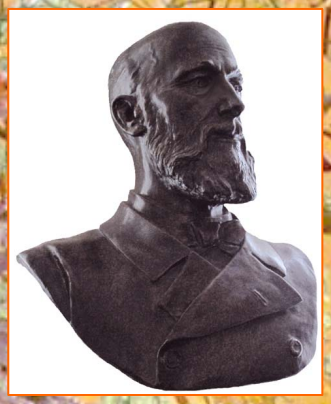
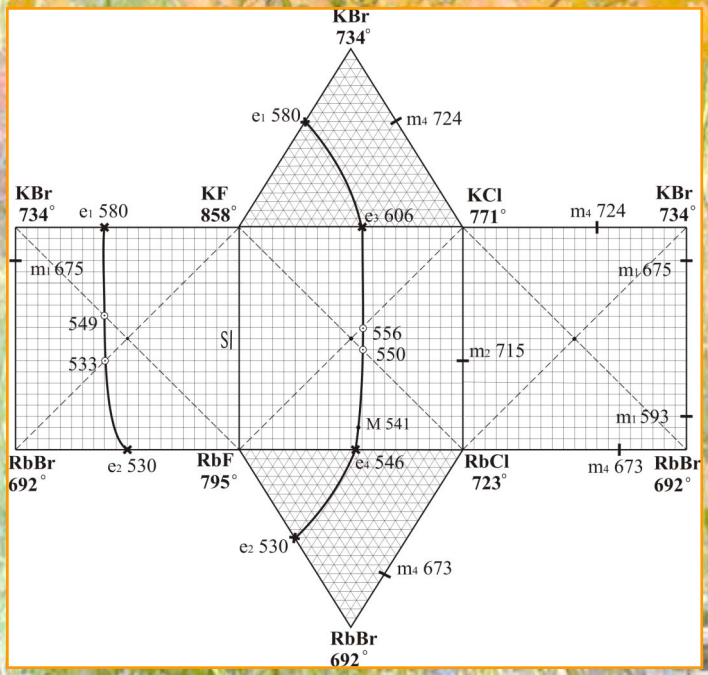
Бутлеровские сообщения

№13, том 27. 2011



ISSN 2074-0212



ISSN 2074-0948

International Edition in English:

Butlerov Communications

Полная исследовательская публикация *Тематический раздел:* Теоретическая и компьютерная химия.
Регистрационный код публикации: 11-27-13-36 *Подраздел:* Математические алгоритмы в химии.
Публикация доступна для обсуждения в рамках функционирования постоянно
действующей интернет-конференции "*Бутлеровские чтения*". http://butlerov.com/readings/
УДК 544.354.081.7:004.021. Поступила в редакцию 8 февраля 2010 г.

Разведочный анализ в термодинамике равновесий. Классификация и прогнозирование силы уксусной кислоты.

© Бондарев Сергей Николаевич,[1] **Зайцева Инна Сергеевна**[2]
и **Бондарев Николай Васильевич**[1]*[+]
[1] *Харьковский национальный университет им. В.Н. Каразина. пл. Свободы, 4. г. Харьков, 61077. Украина. E-mail: n_bondarev@ukr.net*
[2] *Харьковская национальная академия городского хозяйства. Ул. Маршала Бажанова, 17. г. Харьков, 61002. Украина. E-mail: inna.zaitseva123@yandex.ru*

*Ведущий направление; [+]Поддерживающий переписку

Ключевые слова: *разведочные методы, факторний анализ, кластерний анализ, дискриминантний анализ, канонический анализ, деревья классификации, массив данных, константа диссоциации.*

Аннотация

Многомерные разведочные методы – факторный анализ, кластерный анализ, дискриминантный анализ, канонический анализ, построение деревьев классификации применены для классификации и прогнозирования силы уксусной кислоты на основе физико-химических свойств водно-органических растворителей. Построены решающие правила классификации и показана перспективность применения нестатистических методов для классификации и прогнозирования силы уксусной кислоты в водно-органических растворителях.

Введение

Наличие в физикохимии растворов больших массивов экспериментальных данных обусловливает применение математических методов, которые позволяли бы перед проведением углубленного многомерного статистического анализа структурировать информацию, находить объективные закономерности в больших объемах данных, выявлять прогностические возможности обобщений и правил.

Среди таких методов важную роль играют методы многомерного разведочного анализа (РАД) [1]. РАД применяется для нахождения связей между переменными в ситуациях, когда отсутствуют (или недостаточны) априорные представления о природе этих связей.

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 закономерностей в многомерных данных – факторный анализ, кластерный анализ (древовидная классификация, метод *k*-средних), дискриминантный анализ, канонический анализ, построение деревьев классификации [2].

Чтобы выявить влияние свойств растворителя на химическое равновесие (в частности, на константу диссоциации кислоты), необходимо иметь информацию о том, как связаны константа диссоциации и свойства растворителя. По этой информации можно построить решающее правило, которое будет ставить в соответствие произвольному значению константы диссоциации соответствующие значения свойств растворителя.

В работе рассмотрено несколько путей анализа массива данных по константам диссоциации уксусной кислоты и свойствам водно-органических растворителей для выявления имеющихся в них закономерностей и связей.

Наиболее важным итогом применения методов разведочного анализа, является разделение констант диссоциации кислоты на группы (классификация силы кислоты) и выделение значимых свойств (свойства) водно-органических растворителей, лежащих в основе класси-

фикации (дискриминации) и прогнозирования силы уксусной кислоты в еще не изученных средах.

Цель работы: на основе многомерного разведочного анализа свободных энергий Гиббса (констант) диссоциации уксусной кислоты и свойств водно-органических растворителей построить решающее правило классификации и прогнозирования силы уксусной кислоты.

Экспериментальная часть

Особенности разведочного анализа данных (РАД). В факторном анализе проверяются все возможные варианты взаимосвязей между переменными, которые не разделяются на независимые и зависимые. С математической точки зрения факторная модель имеет вид [3]:

$$X_i = A_{i1}F_1 + A_{i2}F_2 + A_{i3}F_3 + \ldots + A_{im}F_m + V_iU_i,$$

где X_i – *i-я* нормированная переменная; A_{im} – нормированный коэффициент множественной регрессии переменной *i* по общему фактору *m*; F_i – общий фактор; V_i - нормированный коэффициент регрессии переменной *i* по характерному фактору *i*; U_i – характерный фактор для переменной *i*; *m* – число общих факторов.

Характерные факторы не коррелируют между собой и с общими факторами. Общие факторы можно выразить линейными комбинациями наблюдаемых переменных:

$$F_i = W_{i1}X_1 + W_{i2}X_2 + W_{i3}X_3 + \ldots + W_{ik}X_k,$$

где F_i – оценка *i-го* фактора; W_i – весовой коэффициент или коэффициент значения фактора; *k* – число переменных.

Веса подбираются так, чтобы первый коэффициент значения фактора объяснял наибольшую долю полной дисперсии. Затем отбирается второй набор весов так, чтобы второй фактор вносил наибольший вклад в остаточную дисперсию при условии, что он не коррелирует с первым фактором.

С помощью кластерного анализа, как и в случае факторного, проверяется весь набор взаимозависимых связей. Цель кластерного анализа – разделение объектов на относительно гомогенные (однородные) группы, исходя из рассматриваемого набора переменных [4]. При использовании кластерного анализа для классификации объектов, он становится составной частью факторного анализа, так как уменьшает число объектов, а не число переменных, сгруппировывая их в меньшее число кластеров.

В работе реализованы иерархические методы кластерного анализа: древовидная кластеризация (агломеративный метод) и дивизивный метод *k*-средних.

Агломеративные методы позволяют строить классификацию в ходе иерархического процесса объединения кластеров. В случае дивизивной стратегии указывается число кластеров, на которое желательно разбить множество.

Таким образом, если *агломеративные* методы идут от объектов при первоначальном отсутствии классов, то в дивизивных в начале процедуры (при $k = 1$) все объекты принадлежат одному кластеру, а затем этот всеобъемлющий кластер разрезается на последовательно уменьшающиеся «части».

Дискриминантный анализ [5] использовался для анализа данных, когда зависимая переменная энергии Гиббса (константы) диссоциации кислоты имела статус категориальной (умеренно слабая, слабая, очень слабая кислота), а предикторы (независимые переменные) были интервальными.

Модель (дискриминантная функция) дискриминантного анализа имеет следующий вид:

$$D = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k,$$

где D – дискриминантный показатель (дискриминант), *b* – дискриминантный коэффициент или вес, X – предиктор или независимая переменная.

С помощью канонического анализа (корреляции) оценивалась степень связи между дискриминантными показателями и группами (кластерами).

Таким образом, кластерный, дискриминантный и канонический анализ наиболее полно отражают особенности многомерного разведочного анализа в классификации объектов, а факторный – в исследовании взаимосвязи между параметрами. Основное назначение построения деревьев классификации (деревьев решений) заключается в подведении итогов разведочного анализа, в структурировании данных и построении решающего правила классификации (принятия решений), обладающего высоким прогностическим потенциалом [6, 7].

Результаты и их обсуждение

Анализируемые данные: зависимые переменные – аквамоляльные [8] стандартные энергии Гиббса ($\Delta G^{o}_{d,HAc}$) диссоциации уксусной кислоты [9-11]; независимые переменные – физико-химические свойства водно-органических растворителей (вода-метанол, вода-этанол, вода-пропан-2-ол) – диэлектрическая проницаемость, электроноакцепторные параметры Димрота-Райхардта, электронодонорные параметром Камлета-Тафта и плотность энергии когезии [12-15].

На первом этапе в качестве метода редукции данных был использован факторный анализ, позволяющий установить число и значимость свойств водно-органических растворителей, влияющих на силу уксусной кислоты.

Целесообразность выполнения факторного анализа определяется наличием корреляций между переменными. Если корреляции между всеми переменными небольшие, то факторный анализ бесполезен. В табл. 1 приведена корреляционная матрица переменных.

Табл. 1. Корреляционная матрица для переменных

Корреляции N = 33					
	$1/\varepsilon^{N}$	E_T^N	B_{KT}	δ^2_N	ΔG^o HAc
$1/\varepsilon^{N}$	1.00	-0.90	0.92	-0.79	0.86
E_T^N	-0.90	1.00	-0.98	0.69	-0.74
B_{KT}	0.92	-0.98	1.00	-0.79	0.82
δ^2_N	-0.79	0.69	-0.79	1.00	-0.94
ΔG^o HAc	0.86	-0.74	0.82	-0.94	1.00

Парные общие корреляции в этой таблице результатов имеют как положительные, так и отрицательные значения. Переменные демонстрируют достаточно высокий уровень кореляции.

Например, переменные E_T^N и B_{KT} коррелированны на уровне 0.98, между E_T^N и $1/\varepsilon^{N}$ коэффициент корреляции -0.9. Таким образом, следует ожидать, что переменные, тесно взаимосвязанные между собой, будут также тесно коррелировать с одним и тем же фактором или факторами.

Для проверки целесообразности использования факторной модели анализа зависимости переменных использован критерий сферичности Бартлетта и критерий адекватности выборки Кайзера-Мейера-Олкина (КМО). Проверка с помощью критерия сферичности основана на преобразовании детерминанта корреляционной матрицы в статистику Хи-квадрат.

Выявлено, что нулевая гипотеза о том, что корреляционная матрица совокупности является единичной матрицей, отклоняется (переменные в генеральной совокупности не коррелируют между собой): значение статистики Хи-квадрат – 290.23; число степеней свободы – 10; значимость – 0.000. Критерий адекватности выборки КМО равен 0.754. Данный коэффициент сравнивает значения наблюдаемых коэффициентов корреляции со значениями частных коэффициентов корреляции. Таким образом, факторный анализ можно рассматривать как приемлемый метод для анализа корреляционной матрицы табл. 1.

Факторы были выделены методом главных компонент (табл. 2). Собственное значение фактора определяет полную дисперсию, присущую данному фактору. Полная дисперсия для всех четырех факторов равна 4, то есть числу переменных. Тогда, дисперсия, обусловленная влиянием первого фактора, равна 3.54 или 88.61% от полной дисперсии (3.54 /4). Второй фактор включает в себя 8.57% дисперсии. Остальные факторы объясняют менее 3% общей дисперсии.

Табл. 2. Результаты факторного анализа

Собственные значения	% общей дисперсии	Кумулятивные собственные значения	Кумулятивная дисперсия %
3.54	88.61	3.54	88.61
0.34	8.57	3.89	97.18
0.11	2.67	3.99	99.85
0.006	0.15	4.00	100.0

Информация о том, сколько общей дисперсии объясняет каждый фактор (табл. 2) позволяет оставить для анализа значимые факторы. На графике (рис. 1) изображены собственные

значения факторов, представленные в табл. 2. Число значимых факторов было выбрано равным двум по критерию *каменистой осыпи* – по максимальному замедлению убывания собственных значений слева направо [2].

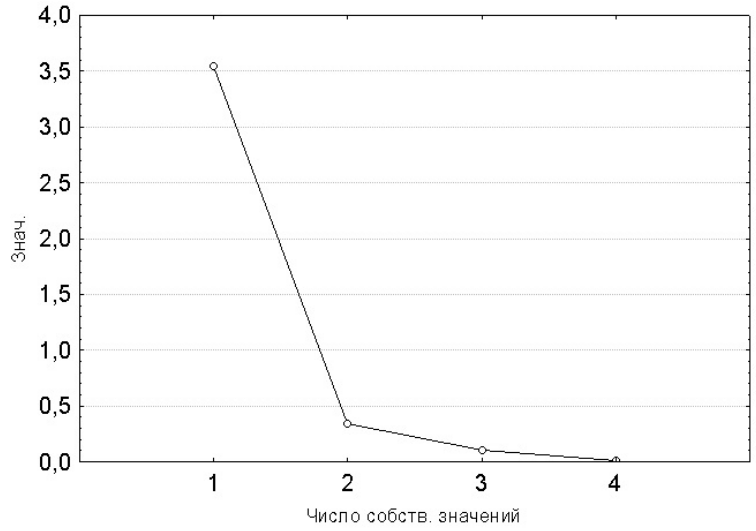


Рис. 1. График каменистой осыпи для выбора значимых факторов

Факторные нагрузки, интерпретируемые как корреляции между факторами и переменными, были выбраны методом вращения варимакс [4], то есть созданием условий, когда факторы отмечены высокими нагрузками на одни переменные и низкими нагрузками на другие переменные (табл. 3). В табл. 3 приведены коэффициенты, используемые для выражения нормированных переменных через латентные (скрытые) факторы.

Табл. 3. Факторные нагрузки на переменные

Факторные нагрузки (Варимакс исходных) Выделение: Главные компоненты (Отмечены нагрузки > 0.70)		
	Фактор 1	Фактор 2
$1/\varepsilon^N$	-0.78	-0.56
E_T^N	0.93	0.36
B_{KT}	-0.86	-0.49
δ^2_N	0.39	0.90
ΔG^o_d HAc	-0.47	-0.87
Общая дисперсия	2.59	2.25
Доля общей дисперсии	0.52	0.45

Как видно из табл. 3, *Фактор 1*, имеет наивысшую нагрузку для переменной E_T^N и наименьшую нагрузку – для переменной δ^2_N. Соответственно, *Фактор 2* имеет наивысшую нагрузку для переменной δ^2_N, самую маленькую нагрузку – для E_T^N, средние нагрузки для остальных переменных.

Из диаграммы рассеяния двух факторов (рис. 2) вытекает, что процедура редукции данных позволила выделить два значимых фактора, влияющих на силу уксусной кислоты в водно-органических растворителях *Фактор 1* и *Фактор 2* и сократить число независимых переменных с четырех до двух.

На концах оси абсцисс расположены переменные $1/\varepsilon^N$, E_T^N и B_{KT}, которые имеют большие нагрузки *Фактором 1*, а в конце оси ординат расположена переменная δ^2_N, имеющая большую нагрузку от *Фактора 2*. Переменные ΔG^o_{dHAc} и E_T^N расположены вдали от осей координат, поэтому они связаны с обоими факторами.

Решением задачи *кластерного анализа* явилось разбиение массива данных по константам диссоциации уксусной кислоты на классы, удовлетворяющее критерию оптимальности.

В качестве критерия качества разбиения на классы в работе выбрана целевая функция W, равная внутригрупповой сумме квадратов отклонений:

$$W = \sum_j (X_j - \bar{X})^2 = \sum_j X_j^2 - (1/n)\left(\sum_j X_j\right)^2,$$

где X_j – вектор независимых переменных j-й константы (энергии Гиббса) диссоциации;

$\bar{X}$ – средний вектор измерений (независимых переменных – свойств растворителя); $j = 1, \ldots, n$.

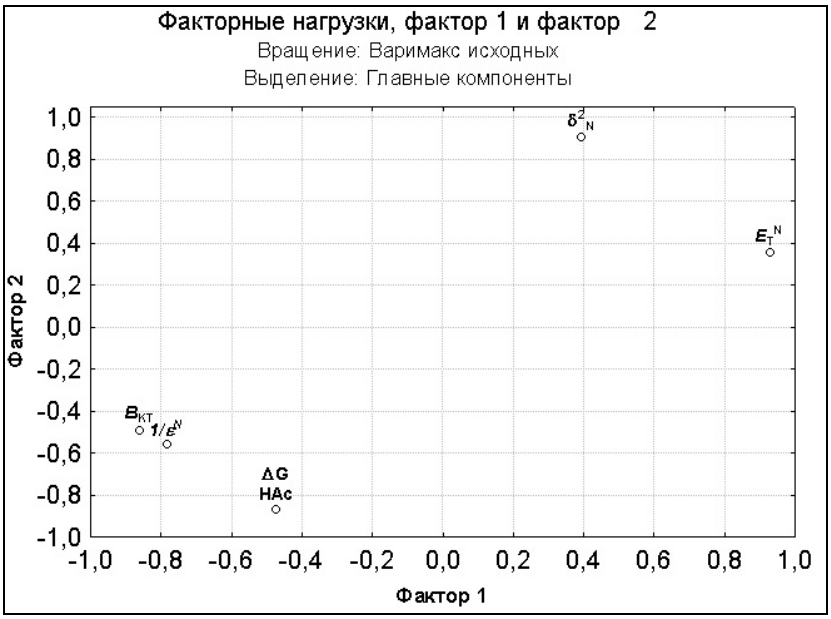


Рис. 2. Диаграмма решения, основанного на вращении двух факторов

Сходство между константами равновесия определялось расстояниями между векторами независимых переменных X_j, X_i; чем меньше расстояние между константами, тем они более схожи.

В качестве функции расстояния $d(X_i, X_j)$ была выбрана евклидова метрика (евклидово расстояние)

$$d_e(X_i, X_j) = \left(\sum_k \left(x_{ik} - x_{jk} \right)^2 \right)^{1/2}$$

где $X_n = x_{n1}, x_{n2}, \ldots, x_{nk}$. $d(X_i, X_j)$ – геометрическое расстояние в многомерном пространстве.

Древовидная кластеризация проведена по методу Варда (правило иерархического объединения в кластеры), в котором в качестве целевой функции используется внутригрупповая сумма квадратов расстояний между каждой константой равновесия и средней константой по кластеру, содержащему эту константу. На каждом шаге объединяются такие два кластера, которые приводят к минимальному увеличению целевой функции, т.е. внутригрупповой суммы квадратов отклонений.

На рис. 3 приведена дендрограмма для изменения энергии Гиббса диссоциации уксусной кислоты в водно-органических растворителях. Пример соответствует случаю 33 наблюдений и 4 характеристикам (независимым признакам) для каждого наблюдения.

На евклидовом расстоянии $d_e(X_i, X_j) = 5$ константы объединены в три кластера, характеризующих силу кислоты: 1 – очень слабая (в растворителях с высоким содержанием неводного компонента – 9 констант диссоциации); 2 – слабая (в растворителях промежуточного состава – 10 констант диссоциации); 3 – умеренно слабая (в растворителях с высоким содержанием воды – 14 констант диссоциации). На расстоянии $d_e(X_i, X_j) = 10$ имеем два кластера: 1 и 2, 3. Окончательно все константы группируются в один кластер на расстоянии $d_e(X_i, X_j) > 21$.

Дивизивный метод k-средних. Вначале кластеры выбирались случайно, а затем изменялась принадлежность констант равновесия к ним, чтобы минимизировать изменчивость внутри кластеров и максимизировать изменчивость между кластерами. Алгоритм случайным образом в пространстве назначает центры будущих кластеров. Затем вычисляет расстояние между центрами кластеров и каждой константой, и константа приписывается к тому кластеру, к которому она ближе всего. Завершив приписывание, алгоритм вычисляет средние значения

для каждого кластера. Процесс повторяется до тех пор, пока центры тяжести не перестают "мигрировать" в пространстве.

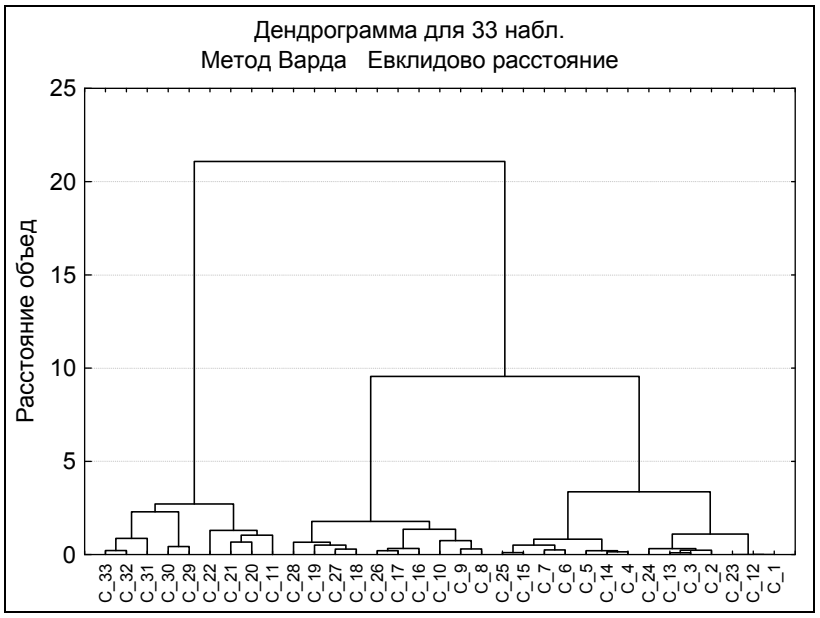


Рис. 3. Дендрограмма – результат последовательной кластеризации энергии Гиббса (констант) диссоциации уксусной кислоты в водно-органических растворителях

Табл. 4. Результаты дисперсионного анализа *k*-средних

	Между SS	cc	Внутри SS	cc	*F*	Уровень значимости
$1/\varepsilon^N$	23.15	2	5.06	30	68.70	0.00
E_T^N	0.29	2	0.19	30	22.37	0.00
B_{KT}	0.82	2	0.37	30	32.84	0.00
δ^2_N	1.14	2	0.42	30	40.42	0.00
ΔG^o_d HAc	26.61	2	5.39	30	74.00	0.00

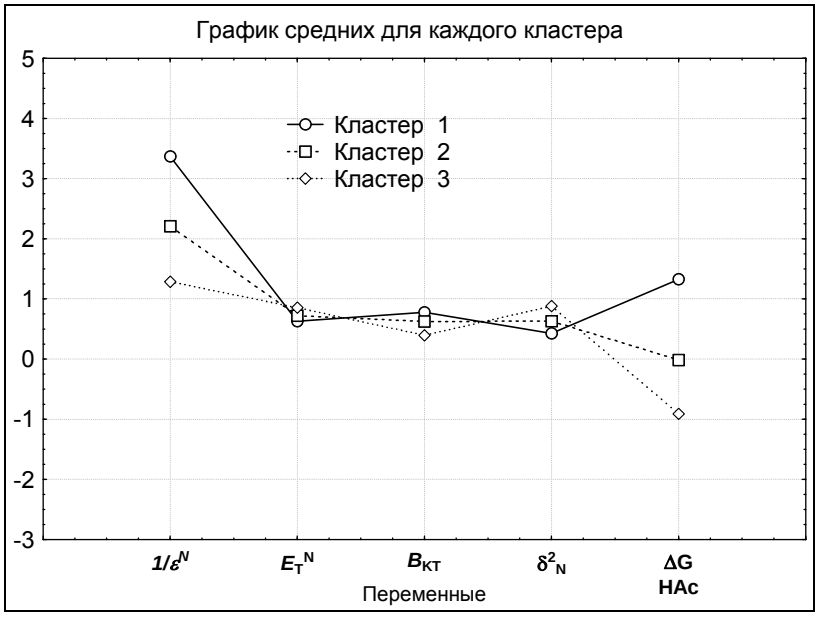


Рис. 4. Средние значения переменных для трех кластеров

На рис. 4 представлены средние значения переменных для трех кластеров. Из графиков, приведенных на рис. 4, видно, что во всех трех кластерах средние значения параметров E_T^N, B_{KT} и δ^2_N незначительно отличаются друг от друга. Таким образом, можно заключить, что

определяющим фактором распределения констант диссоциации уксусной кислоты по кластерам является диэлектрическая проницаемость водно-органических растворителей.

В табл. 4 приведены статистически значимые (наблюдаемый критерий Фишера $F > 22$, уровень значимости практически равен нулю) результаты дисперсионного анализа *k*-средних – значения межгрупповых (Между SS) и внутригрупповых (Внутри SS) дисперсий зависимой и независимых переменных при соответствующих степенях свободы (сс).

Чем меньше значение внутригрупповой дисперсии и больше значение межгрупповой дисперсии, тем лучше независимая переменная характеризует принадлежность константы диссоциации к кластеру. Как видно из табл. 4 в наибольшей мере этому условию отвечает диэлектрическая проницаемость.

Распределение констант диссоциации кислоты по трем кластерам в соответствии с евклидовым расстоянием до центра кластера (табл. 5) практически совпадает с древовидной кластеризацией по методу Варда (рис. 3) на евклидовых расстояниях объединения в кластеры близких 5.

Табл. 5. Энергии Гиббса (константы) диссоциации уксусной кислоты в кластерах и расстояния до центра кластера

Элементы кластера номер 1 и расстояния до центра кластера. Кластер содержит 9 набл. Интервал изменения p*K*: 7.025-9.907									
	C-11	C-20	C-21	C-22	C-29	C-30	C-31	C-32	C-33
Расстояние	0.52	0.35	0.17	0.43	0.48	0.35	0.22	0.33	0.42

Элементы кластера номер 2 и расстояния до центра кластера. Кластер содержит 11 набл. Интервал изменения p*K*: 5.945-7.583											
	C-7	C-8	C-9	C-10	C-16	C-17	C-18	C-19	C-26	C-27	C-28
Расстояние	0.27	0.18	0.16	0.32	0.25	0.12	0.08	0.24	0.18	0.18	0.38

Элементы кластера номер 3 и расстояния до центра кластера. Кластер содержит 13 набл. Интервал изменения p*K*: 4.756-5.706													
	C-1	C-2	C-3	C-4	C-5	C-6	C-12	C-13	C-14	C-15	C-23	C-24	C-25
Расстояние	0.24	0.14	0.05	0.06	0.15	0.25	0.24	0.07	0.10	0.23	0.24	0.07	0.23

Задача *дискриминантного анализа* состояла: в определении дискриминантных функций (линейных комбинаций независимых переменных), которые наилучшим образом различают (дискриминируют) классы (кластеры) зависимой переменной (константы диссоциации уксусной кислоты); в проверке существования между группами значимых различий посредством независимых переменных; в определении предикторов, вносящих наибольший вклад в межгрупповые различия констант диссоциации; в отнесении случаев к одной из групп (классификация силы кислоты), исходя из значений предикторов; в оценке точности разделения констант диссоциации на группы.

При пошаговом анализе алгоритм отбирал переменные, дающие наиболее значимый (дополнительный) вклад в дискриминацию между совокупностями (классами). В табл. 6 приведены финальные результаты дискриминантного анализа.

В первом столбце таблицы приведены значения λ-Уилкса, являющиеся результатом исключения соответствующей переменной из модели. Чем больше значение λ, тем более желательно присутствие этой переменной в процедуре дискриминации.

Частная лямбда Уилкса характеризует единоличный вклад соответствующей переменной в дискриминационную силу модели. Чем меньше значение частной λ, тем больше вклад переменной в общую дискриминацию. Из таблицы видно, что переменная $1/\varepsilon^N$ вносит наибольший вклад в дискриминацию констант диссоциации кислоты. Об этом же свидетельствуют и частная статистика λ-Уилкса.

Для дальнейшего исследования природы дискриминации проведен *канонический анализ* на основе оцениваемых дискриминантных функций и с учетом того, что каждая последующая дискриминантная функция будет вносить все меньший и меньший вклад в общую дискри-

минацию. Максимальное число оцениваемых функций может быть равным числу переменных. В работе оценивались две дискриминантные функции.

Табл. 6. Итоги дискриминантного анализа

Итоги анализа дискриминантных функций; Переменных в модели: 4; Группирующая: ΔG^{o}_{d} HAc (3 гр.) Лямбда Уилкса: 0.10; F (8.54) = 14.453 p < 0.0000					
	λ-Уилкса	Частная λ-Уилкса	F-исключения	p-уровень	Толерантность
$1/\varepsilon^N$	0.20	0.50	13.63	0.00	0.38
E_T^N	0.10	0.98	0.26	0.77	0.03
B_{KT}	0.10	0.99	0.18	0.83	0.03
δ^2_N	0.11	0.90	1.49	0.24	0.45

В табл. 7 приведены пошаговые значения Хи-квадрат критерия для канонических корней – дискриминантных функций. В первой строке приведен критерий значимости для двух дискриминантных функций (для всех корней). Вторая строка содержит значимость второй дискриминантной функции после удаления первой (первого корня). Как видно из таблицы, статистически значима только первая дискриминантная функция, для второй очень низкое значение имеют критерий Хи-квадрат и уровень значимости.

Табл. 7. Критерий Хи-квадрат с последовательно исключаемыми корнями

Удаленные корни	Собственные значения	Канонические значения	λ-Уилкса	Хи-квад.	степени свободы	p-уровень
0	7.49	0.94	0.10	65.24	8	0.00
1	0.16	0.37	0.86	4.28	3	0.23

Наибольший вклад в дискриминантную функцию 1 вносит обратная величина диэлектрической проницаемости и E_T^N (табл. 8). В таблице приведено собственное значение дискриминантной функции 1 (7.49) и кумулятивная доля объясненной дисперсии, накопленной каждой функцией. Таким образом, функция 1 ответственна за 97.88% объясненной дисперсии то есть 97.88% всей дискриминирующей мощности определяется этой функцией.

Табл. 8. Стандартизованные коэффициенты дискриминантных функций

Стандартизованные коэффициенты для канонических переменных		
Переменные	Корень 1	Корень 2
$1/\varepsilon^N$	1.15	1.06
E_T^N	0.80	-0.11
B_{KT}	0.30	-1.50
δ^2_N	-0.50	0.14
Собственные значения	7.49	0.16
Кумулятивная доля	0.98	1.00

В табл. 9 приведены объединенные внутригрупповые корреляции переменных с соответствующими дискриминантными функциями – структурные коэффициенты. Эти коэффициенты используют для содержательной интерпретации функций, в отличие от коэффициентов дискриминантной функции (табл. 8), которые отражают вклад каждой переменной в функцию. У переменных $1/\varepsilon^N$ и δ^2_N наибольшие корреляции с дискриминантной функцией 1, у переменных E_T^N и B_{KT} – наибольшие корреляции с дискриминантной функцией 2, но, как отмечено ранее, эта функция статистически незначима.

Дискриминантная функция 1 хорошо идентифицирует константы диссоциации уксусной кислоты, относящиеся к первому и третьему классам (значения среднего по модулю близкие), а дискриминантная функция 2 – практически одинаково идентифицирует все три класса

констант диссоциации кислоты (табл. 10). Но дискриминантная функция 2 определяет лишь 2.12 % (табл. 8) дискриминирующей мощности (100-97.88%).

Табл. 9. Корреляции переменных и дискриминантных функций

Матрица факторной структуры Корреляции переменных и дискриминантных функций (объединенные внутригрупповые корреляции)		
	Корень 1	Корень 2
$1/\varepsilon^N$	0.78	0.02
E_T^N	-0.44	0.51
B_{KT}	0.53	-0.64
δ^2_N	-0.60	0.50

Табл. 10. Средние значения канонических переменных

Средние канонических переменных		
	Корень 1	Корень 2
G_1:1	3.76	0.30
G_2:2	0.16	-0.54
G_3:3	-2.74	0.25

Константы диссоциации уксусной кислоты, принадлежащие одинаковым группам (характеризующим силу кислоты), локализованы в определенных областях плоскости (рис. 5). При этом расстояние между центроидами групп 1 и 3 намного больше, чем расстояния между центроидами групп 1 и 2, 2 и 3, что свидетельствует о том, что сила кислоты в растворителях с большим содержанием воды значительно отличается от силы кислоты в растворителях с большим содержанием спирта. Из диаграммы также следует, что дискриминация по дискриминантной функции 1 более выражена, чем по дискриминантной функции 2.

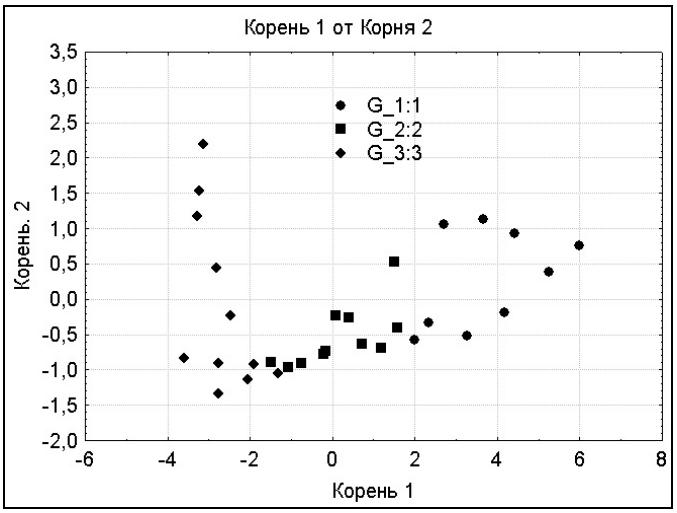


Рис. 5. Диаграмма рассеяния для канонических значений

Коэффициенты функций классификации – линейных функций, которые вычисляются для каждой группы (табл. 11), могут быть использованы для классификации наблюдений (констант диссоциации уксусной кислоты). Наблюдение приписывается той группе, для которой классификационная функция имеет наибольшее значение. Классификационные функции для трех групп констант диссоциации уксусной кислоты в водно-органических растворителях имеют вид:

$$G\_1:1 = -3792.99 + 123.18 \cdot 1/\varepsilon^N + 5872.45 \cdot E_T^N + 4030.93 \cdot B_{KT} + 795.35 \cdot \delta^2_N$$
$$G\_2:2 = -3743.55 + 110.91 \cdot 1/\varepsilon^N + 5837.51 \cdot E_T^N + 4032.59 \cdot B_{KT} + 809.39 \cdot \delta^2_N$$
$$G\_3:3 = -3709.62 + 104.85 \cdot 1/\varepsilon^N + 5807.43 \cdot E_T^N + 4014.23 \cdot B_{KT} + 822.45 \cdot \delta^2_N$$

В табл. 12 представлена информация о количестве и проценте корректно классифицированных констант диссоциации кислоты в каждой группе.

Как следует из табл. 12, не удалось правильно классифицировать две константы диссоциации. Одна константа первого класса классифицирована как константа второго класса

и одна константа третьего класса также классифицирована как константа второго класса. Следовательно, точность классификации силы уксусной кислоты на основе дискриминационного и канонического анализа составляет 91%.

Табл. 11. Коэффициенты классификационных функций

Функции классификации; группировка: ΔG^{o}_{d} HAc			
Переменная	G_1:1 P = 0.27	G_2:2 P = 0.33	G_3:3 P = 0.39
$1/\varepsilon^{N}$	123.18	110.91	104.85
E_{T}^{N}	5872.45	5837.51	5807.43
B_{KT}	4030.93	4032.59	4014.23
δ^{2}_{N}	795.35	809.39	822.45
Константа	-3792.99	-3743.55	-3709.62

Для анализа структуры данных и выведения решающего правила классификации было построено *дерево классификации* (рис. 6). Процесс построения дерева классификации состоял из четырех основных этапов: *а*) выбора критерия точности прогноза; *б*) выбора вариантов ветвления; *в*) определения момента прекращения ветвления; *г*) определение оптимального размера дерева.

Табл. 12. Матрица классификации силы уксусной кислоты

Группа	Процент	Наблюдаемые классы Столбцы		
		G_1:1 P = 0.27	G_2:2 P = 0.33	G_3:3 P = 0.39
		Предсказанные классы		
G_1:1	88.89	8	1	0
G_2:2	90.91	0	10	1
G_3:3	92.31	0	1	12
Всего	90.91	8	12	13

а) Наиболее точный прогноз связан с наименьшей ценой, то есть долей неправильно классифицированных наблюдений. Цена неправильной классификации объектов была выбрана одинаковой – все недиагональные элементы матрицы цен ошибок классификации (прогнозируемые классы – по строкам, наблюдаемые классы – по столбцам) принимались равными 1. Априорные вероятности оценивались пропорционально размерам классов зависимой переменной (константы диссоциации кислоты).

б) Выбран способ ветвления по значениям предикторных переменных. Такие ветвления проводятся последовательно, начиная с корневой вершины, затем переходят к дочерним вершинам, пока дальнейшее ветвление не прекратится и "неразветвленные" вершины окажутся терминальными (или, как их иногда называют, листьями), то есть узлами дерева, начиная с которых ветвление больше не происходит. В работе осуществлен полный перебор вариантов одномерного ветвления методом *CART*. В качестве критерия согласия для выбора наилучшего из всех возможных вариантов ветвления использована мера Джини. Мера Джини однородности вершины принимает нулевое значение, если в данной вершине имеется всего один класс.

в) Остановка ветвления дерева проведена отсечением по ошибке классификации на основе правила стандартной ошибки.

В заголовке графа (рис. 6) приведена общая информация (табл. 13), согласно которой полученное дерево классификации имеет 2 ветвления и 3 терминальные вершины. Началом дерева считается самая верхняя решающая вершина, которую иногда также называют корнем дерева.

Первоначально все 33 константы приписываются к этой корневой вершине и предварительно классифицируют кислоту как умеренно слабая (3 класс) – на это указывает число 3 в правом верхнем углу вершины. Класс 3 был выбран для начальной классификации потому, что число констант, характеризующих кислоту как умеренно слабая (13) немного больше, чем число констант, характеризующих кислоту как слабая (11) и очень слабая (9). В левом верхнем углу графа имеется надпись – легенда, указывающая, какие столбики гистограммы вершины соответствуют силе кислоты: 1 – очень слабая; 2 – слабая; 3 – умеренно слабая.

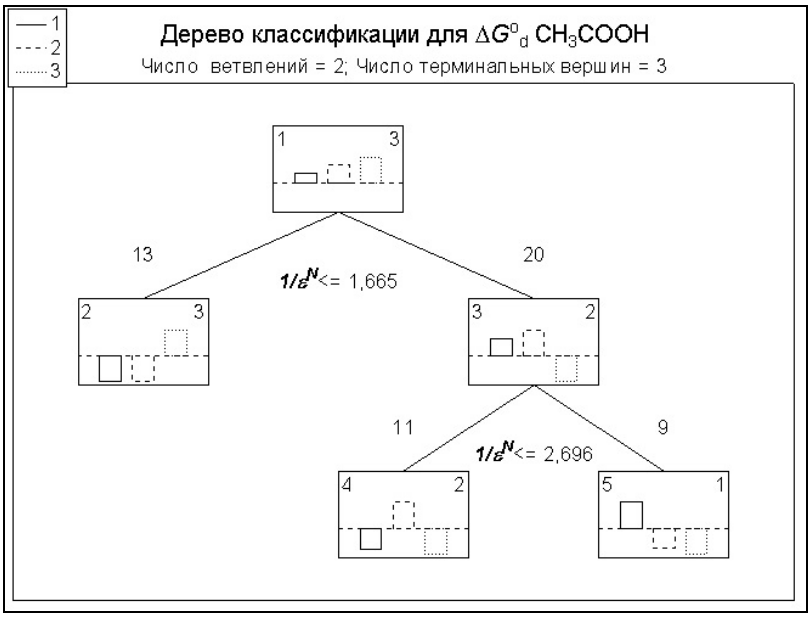


Рис. 6. Дерево классификации силы уксусной кислоты по значениям обратной диэлектрической проницаемости водно-органических растворителей

Табл. 13. Структура дерева классификации, наблюдаемые, предсказанные классы, условия ветвления

Структура. Дочерние вершины, наблюдаемые, предсказанные классы, условия ветвления.								
Вершина	Левая вершина	Правая вершина	Класс 1	Класс 2	Класс 3	Предс-казанный класс	Ветвление по постоянной	Ветвление по переменной
1	2	3	9	11	13	3	-1.665	$1/\varepsilon^N$
2			0	0	13	3		
3	4	5	9	11	0	2	-2.696	$1/\varepsilon^N$
4			1	10	0	2		
5			8	1	0	1		

Корневая вершина разветвляется на две новых вершины. Под корневой вершиной приведено условие данного ветвления. Из него следует, что константы диссоциации, которым соответствуют значения обратной приведенной диэлектрической проницаемости меньшие или равное 1.665, отнесены к вершине номер 2 и классифицированы как константы умеренно слабой кислоты (3), а константы, которым соответствуют значения $1/\varepsilon^N > 1.665$, приписаны к вершине 3 и предположительно классифицированы как константы слабой кислоты (2).

Числа 13 и 20 над вершинами 2 и 3 соответственно обозначают число констант диссоциации, попавших в эти две дочерние вершины из родительской корневой вершины. Затем точно так же разветвляется вершина 2. В результате 11 констант со значениями $1/\varepsilon^N$ меньшими или равными 2.696 приписываются к вершине 4 и классифицируют кислоту как слабая, а остальные 9 констант диссоциации с $1/\varepsilon^N > 2.696$ – к вершине 5 и классифицируют кислоту как очень слабая.

В табл. 14 приведена информация о том, сколько констант каждого из наблюдаемых классов по результатам классификации были ошибочно отнесены к другому классу.

Из табл. 14 видно, что одна константа класса 1 неправильно классифицирована как константа класса 2 и одна константа класса 2 неправильно классифицирована как константа класса 1. Все константы диссоциации уксусной кислоты класса 3 классифицированы верно. Из анализа табл. 12, 14 следует, что результаты построения дерева классификации и дискриминантного анализа практически совпадают.

Таким образом, получено решающее правило, состоящее из двух этапов и позволяющее классифицировать уксусную кислоту как умеренно слабую, слабую или очень слабую по значениям диэлектрической проницаемости водно-органических растворителей.

Табл. 14. Ошибки классификации констант диссоциации уксусной кислоты

Ошибки классификации на обучающей выборке Предсказанные (строки) и наблюдаемые (столбцы) Объем обучающей выборки N = 33			
Класс	Класс 1	Класс 2	Класс 3
1		1	0
2	1		0
3	0	0	

На первом этапе проверяется справедливость неравенства $1/\varepsilon^N \leq 1.665$. Если неравенство выполняется, то константы диссоциации будут классифицировать кислоту как умеренно слабую (pK: 4.76-5.71). Если $1/\varepsilon^N > 1.665$, то константы диссоциации классифицируют кислоту как слабую или очень слабую.

На втором этапе для констант, неклассифицированных на предыдущем этапе как константы диссоциации кислоты умеренной силы, проверяется справедливость неравенства $1/\varepsilon^N \leq 2.696$. Если неравенство выполняется, то константы характеризуют кислоту как слабую (pK: 5.94-7.58). Если неравенство не выполняется, то константы классифицируют кислоту как очень слабая (pK: 7.02-9.91).

Прогностические возможности полученного правила проверены на константах диссоциации уксусной кислоты в водно-диоксановых и водно-диметилсулфоксидных растворителях (табл. 15).

Табл. 15. Подтверждение выполнимости правила классификации силы уксусной кислоты в водно-органических растворителях

мол. д DO	$1/\varepsilon^N$	p$K_{эксп}$	Правило, класс (кластер), интервал изменений pK		
	Вода – диоксан (DO) – уксусная кислота				
0	1	4.76	$1/\varepsilon^N \leq 1.665$	3	4.76 – 5.71
0.1	1.645	5.70	$1/\varepsilon^N \leq 1.665$	3	4.76 – 5.71
0.2	2.625	6.67	$1.665 < 1/\varepsilon^N < 2.696$	2	5.94 – 7.58
0.3	3.848	7.71	$1/\varepsilon^N > 2.696$	1	7.02 – 9.91
0.4	5.326	8.90	$1/\varepsilon^N > 2.696$	1	7.02 – 9.91
0.5	9.229	10.32	$1/\varepsilon^N > 2.696$	1	7.02 – 9.91
мол. д. DMSO	Вода – диметилсульфоксид (DMSO) – уксусная кислота				
0	1	4.76	$1/\varepsilon^N \leq 1.665$	3	4.76 – 5.71
0.1	1.025	5.17	$1/\varepsilon^N \leq 1.665$	3	4.76 – 5.71
0.2	1.047	5.79	$1/\varepsilon^N \leq 1.665$	3	4.76 – 5.71

Заключение

На основе многомерного разведочного анализа построено решающее правило классификации и прогнозирования силы уксусной кислоты по диэлектрической проницаемости водно-органических растворителей. Высокий прогностирующий потенциал правила классификации подтвержден на независимых экспериментальных данных.

Полученное решающее правило классификации позволяет по единственной переменной – диэлектрической проницаемости водно-органического растворителя предсказывать не только класс силы уксусной кислоты, но и значение показателя константы диссоциации в пределах единицы pK.

Результаты разведочного анализа являются необходимым условием первичной обработки массивов данных с целью выявления значимых предикторов и связей между переменными перед применением углубленных статистических методов анализа – множественная линейная или нелинейная регрессия, общие линейные модели, общие регрессионные модели, обобщенные линейные и нелинейные модели, нелинейное оценивание.

Литература

[1] Тьюки Дж. Анализ результатов наблюдений. Разведочный анализ. *М.: Мир*. **1981**. 696с.

[2] Халафян А.А. Статистический анализ данных. STATISTICA 6.0. 2-е изд. испр. и доп. *Краснодар: КубГУ*. **2005**. 308с.

[3] Малхорта Н.К. Маркетинговые исследования. Практическое руководство. *М.: Издательский дом "Вильямс"*. **2002**. 960с.
[4] Наследов А.Д. SPSS: Компьютерный анализ данных в психологии и социальных науках. *СПб.: Питер*. **2005**. 416с.
[5] Ким Дж.-О., Мьюллер Ч.У., Клекка У.Р. Факторный, дискриминантный и кластерный анализ. *М.: Финансы и статистика*. **1989**. 216с.
[6] Боровиков В. STATISTICA. Искусство анализа данных на компьютере: Для профессионалов. 2-е изд. *СПб.: Питер*. **2003**. 686с.
[7] Барсегян А.А., Куприянов М.С., Степаненко В.В. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. *СПб.: БХВ-Петербург*. **2007**. 384с.
[8] E.N. Tsurko, N.V. Bondarev. Mathematical modeling of solvent parameters' relevant contribution to the alpha-amino acid (valine, alpha-alanine) solvation in H_2O–MeOH, H_2O–EtOH and H_2O– PrOH-2. *J. Mol. Liquids*. **2007**. No.131-132. P.151-157.
[9] Никольский Б.П. Справочник химика Том 3 Изд. 2. *Л.: Изд-во «Химия»*. **1965**. 1008с.
[10] Лебедь В.И., Бондарев Н.В. Константы диссоциации и термодинамические характеристики диссоциации уксусной и бензойной кислот в смесях вода – метанол, вода – диоксан. *Журн. физ. химии*. **1982**. Т.56. №1. С.30-33.
[11] Лебедь В.И., Бондарев Н.В., Пауленова А. Константы диссоциации и термодинамические характеристики диссоциации и сольватации уксусной кислоты в смесях вода – пропанол-2. *Журн. физ. химии*. **1987**. Т.61. №6. С.1487-1491.
[12] Зайцева И.С., Ельцов С.В., Кабакова Е.Н., Бондарев Н.В. Корреляционный анализ влияния эффектов среды на энергетику комплексообразования катионов натрия и калия с эфиром 18-краун-6 в водно-органических растворителях. *Журн. общ. химии*. **2003**. Т.73. Вып.7. С.1079-1084.
[13] Афанасьев В.Н., Ефремова Л.С., Волкова Т.В. Физико-химические свойства бинарных растворителей. Водосодержащие системы. *Иваново: ИХР РАН*. **1988**. 412с.
[14] C. Kalidas, G. Hefter, Y. Marcus. Gibbs energies of transfer of cations from water to mixed aqueous organic solvents. *Chem. Rev*. **2000**. Vol.100. No.3. P.819-852.
[15] G. Hefter, Y. Marcus, W.E. Waghorne. Enthalpies of Transfer of Electrolytes and Ions between Water and Mixed Aqueous Solvents. *Chem. Rev*. **2002**. Vol.102. No.8. P.2773-2836.