УДК 542.121

К РАЗРАБОТКЕ НЕЧЕТКИХ КРИТЕРИЕВ ИДЕНТИФИКАЦИИ В КАЧЕСТВЕННОМ ХИМИЧЕСКОМ АНАЛИЗЕ

© 2005 А. В. Пантелеймонов, Ю. В. Холин

Показана возможность построения критериев сходства объектов, идентифицируемых по спектроскопическим данным, с использованием аппарата теории нечетких множеств. Описана процедура вычисления функции принадлежности и проведено сравнение предложенных критериев с аналогами.

Многие современные методы идентификации веществ по результатам спектроскопических или хроматографических измерений предусматривают сравнение многомерных массивов данных, характеризующих «эталон» («образец») и исследуемое вещество. Как правило, такие массивы представляют собой таблично заданные зависимости отклика (откликов) от предикторов, описывающих условия выполнения измерений, необходимых для идентификации. Для принятия решений: «исследуемое вещество совпадает с эталоном», «исследуемое вещество отличается от эталона», «определенное суждение невозможно», исследуют количественные критерии сходства эталона и исследуемого вещества (критерии предусматривают оценку близости откликов эталона и аналита при одинаковых значениях предикторов). До настоящего времени наибольший прогресс в конструировании критериев сходства достигнут В.И. Вершининым [1,2]. Он ввел «суммарную надежность идентификации» аналита

$$P = 1 - \alpha - \beta , \qquad (1)$$

где α – вероятность ложной идентификации вещества если на самом деле его в пробе нет (ошибка первого рода), β – вероятность его пропуска, то есть необнаружения этого вещества в пробе, если оно там присутствует (ошибка второго рода). Постулируя Лапласовское распределение погрешностей, характеризующих условия проведения эксперимента, и Гауссову форму пиков, В.И. Вершинин и соавт. показали [1], как вычислять ошибки α и β для данных ИК- и ЯМР-спектроскопии и хроматографии.

Оценить различие/сходство объектов позволяет вычисление расстояния (d) между ними. Оценить значение расстояния можно различными способами [3, 4], но чаще всего используют Евклидову метрику:

$$d = \sqrt{\sum_i \left(a(x_i) - b(x_i)\right)^2} , \qquad (2)$$

где $a(x_i)$ и $b(x_i)$ – величины откликов измерительной системы для эталона и аналита при аналитической позиции x_i (в случае многомерных откликов необходимо их автомасштабное преобразование). Отметим очевидную аналогию между вычислением Евклидова расстояния между объектами и оцениванием близости модельных значений случайной величины к экспериментальным с помощью статистики χ^2:

$$\chi^2 = \sum_i \left(\frac{y_i - \hat{y}_i}{s_i} \right)^2 , \qquad (3)$$

где y_i и $\hat{y}_i$ – экспериментальное и модельное значения случайной величины y_i, s_i – ее стандартное отклонение, i – номер экспериментальной точки. Если принять, что все y_i измерены с одинаковой относительной погрешностью δ, $s_i = \delta \cdot y_i$.

Как при оценке суммарной надежности идентификации, так и при оценке степени сходства между объектами на основе вычисления расстояний, исследователь принимает решение об идентичности аналита и эталона или о принадлежности аналита определенному классу веществ, руководствуясь критическими (граничными) значениями критериев, которые рассчитывают на основе предположений об известном (включая и величины параметров) законе распределения экспериментальных погрешностей.

Трудности анализа состоят, главным образом, в невозможности абсолютно точно контролировать условия измерений и в отсутствии априорной информации о распределении экспериментальных погрешностей. Кроме того, критерий «суммарная надежность идентификации» не

является универсальным (например, процедура его применения для обработки данных оптических методов не разработана). Предложить один универсальный критерий сходства объектов невозможно, но особый интерес вызывают робастные критерии, не слишком чувствительные к гипотезам о распределении погрешностей.

В настоящей работе для количественной оценки сходства аналита и эталона привлечен аппарат теории нечетких множеств, обеспечивающий робастность оценивания. Теория нечетких множеств [5] рассматривает числа или отношения как не имеющие четких границ. Рис. 1 иллюстрирует справедливость высказывания $x > y$ при интерпретации этого отношения обычным множеством (а) и с использованием нечетких множеств (б). В последнем случае легко оценить и степень истинности высказывания $x >> y$ (показана градиентом цвета).

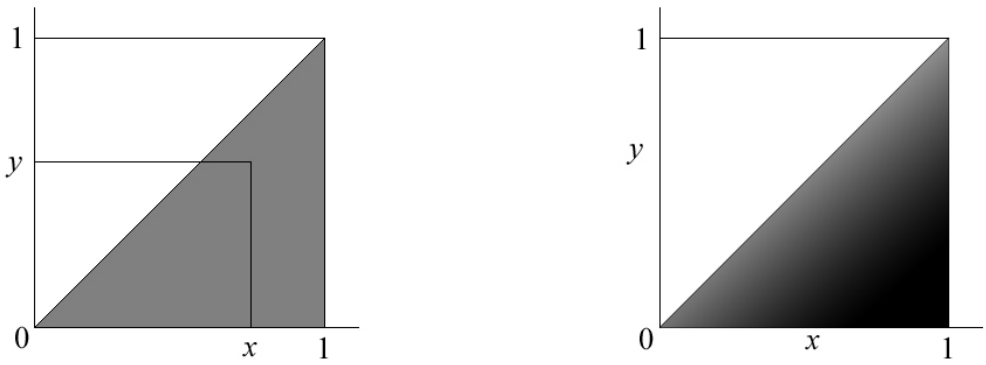


Рис. 1. Отношение x > y в обычных (а) и нечетких (б) множествах.

Представление результатов измерений в виде нечетких чисел производится с помощью процедуры «размывания» («fuzzyfication») обычных чисел. Простейший способ представления обычного числа в виде нечеткого иллюстрирует рис. 2.

Здесь x – обычное число (результат измерения), x_L и x_R – границы нечеткости числа, μ – вводимая в теории нечетких множеств функция принадлежности, являющаяся мерой истинности высказывания «результат измерения равен числу x» ($0 \le \mu \le 1$). Степень близости нечетких чисел оценивают с помощью операции пересечения [6]. На рис. 3 нечеткие числа x_1 и x_2 принадлежат одному множеству с $\mu = 0.83$.

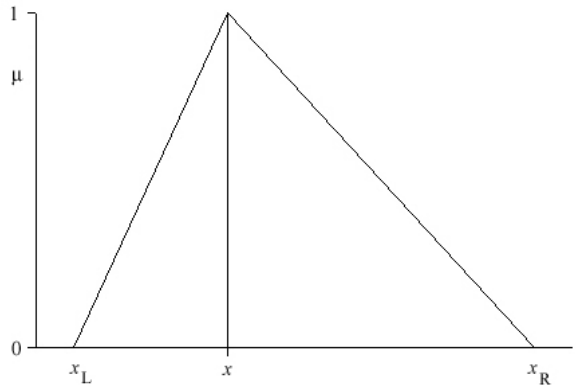


Рис. 2. Представление числа x в виде нечеткого.

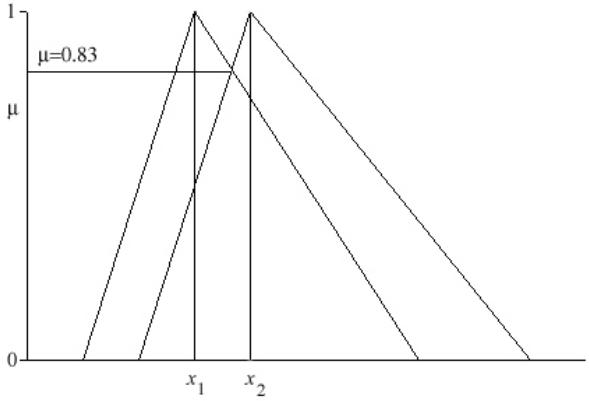


Рис. 3. Оценка степени принадлежности двух нечетких чисел одному множеству.

Оценивая сходство аналита с эталоном, соответствующие массивы результатов измерений можно представить в виде нечетких чисел. После для каждой из пар откликов (i), полученных для одного и того же предиктора, найти функции принадлежности одному множеству (μ_i), и

для набора функций принадлежности вычислить суммарную мощность так же, как для обычного множества [6]:

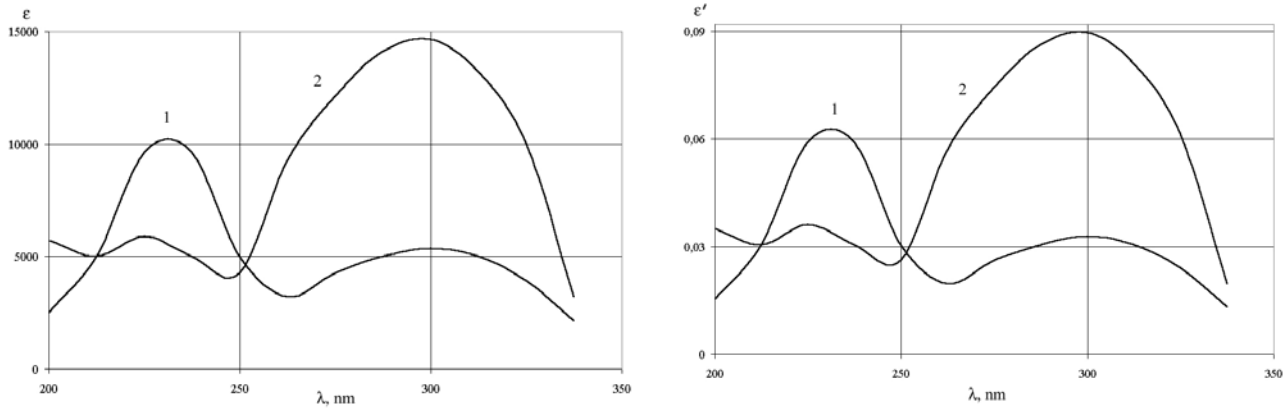
$$\mu_{\text{сум}} = \text{card}(P) = \frac{1}{N}\sum_{i=1}^{N}\mu_i \,, \qquad (4)$$

где N – число измерений, card(P) – мощность множества P функций принадлежности μ_i. Мощность $\mu_{\text{сум}}$ выступает критерием сходства аналита с эталоном. Поскольку теория нечетких множеств оперирует с субъективными вероятностями, не лишено субъективности и решение об идентичности эталона и аналита, принимаемое при высоких значениях $\mu_{\text{сум}}$ (скажем, при $\mu_{\text{сум}} > 0.95$), или же вывод об их полном различии (например, при задании границы $\mu_{\text{сум}} < 0.3$). Вместе с тем, явное введение «субъективного критерия» обнажает неопределенность процедур принятия решений в качественном химическом анализе и на данном этапе исследований не может рассматриваться как препятствие для исследования свойств критерия $\mu_{\text{сум}}$ в сравнении с имеющимися аналогами.

Понятие «нечеткое число» и соответствующие операции естественным образом обобщаются и на многомерные объекты. Это особенно важно, если учесть, что не всегда можно обеспечить идентичность условий измерений для аналита и эталона. В таком случае процедуре «размывания» целесообразно подвергнуть не только отклики, но и предикторы.

Возможности развиваемого подхода проверены при оценке сходства УФ-спектров бензоилацетона, измеренных в разных растворителях. Зависимости молярных коэффициентов поглощения (ε, л·моль$^{-1}$·см$^{-1}$ – отклики) от длин волн (λ, нм – предикторы) заимствованы из работы [8] (рис. 4а). Спектры были представлены в виде таблиц шагом по длинам волн 10 нм. Так как значение критерия d варьируется в пределах $0 \le d < \infty$ [3], спектры предварительно нормировали на общую площадь с целью повышения информативности критерия. Евклидово расстояние между нормированными спектрами $d = 0.20$ (нормированные спектры представлены на рис. 4б). Значение критерия Евклидова расстояния, вычисленное без процедуры автомасштабного преобразования, составляет $d = 2\cdot10^4$.

Вычисляя критерий $\mu_{\text{сум}}$, выполнили процедуру «размывания» и для откликов, и для предикторов. Необходимо отметить, что величины границ нечеткости являются параметрами, выбираемыми пользователем произвольно в зависимости от точности выполнения эксперимента, условий его проведения и природы сравниваемых веществ (в случае задачи обнаружения токсикантов, наркотических или взрывчатых веществ, значения параметров нечеткости можно увеличить). Изменение границ нечеткости и откликов, и предикторов не слишком сильно сказывается на величине $\mu_{\text{сум}}$ (табл. 1, 2), что служит аргументом в пользу использования развиваемого подхода, подтверждая его робастность. Разумеется, низкие значения $\mu_{\text{сум}}$ исключают вывод об идентичности спектров бензоилацетона в воде и в эфире.



а б

Рис. 4. Измеренные (а) и нормированные (б) спектры бензоилацетона в воде (1) и эфире (2).

Таблица 1. Величины критерия $\mu_{сум}$ при различных величинах границ нечеткости ε (границы нечеткости длин волн $\lambda_L = \lambda_R = 10$ нм)

ε_L	ε_R	$\mu_{сум}$
1000	1000	0.458
2000	2000	0.487
3000	3000	0.527

Таблица 2. Величины критерия $\mu_{сум}$ при различных величинах границ нечеткости длин волн (границы нечеткости интенсивностей $\varepsilon_L = \varepsilon_R = 1000$)

λ_L	λ_R	μ
10	10	0.458
20	20	0.523
30	30	0.544

Второй пример иллюстрирует полезность критерия $\mu_{сум}$ для принятия решений и в более сложных ситуациях. На рис. 5 представлены спектры поглощения 2-бензолсульфонимидо-5-метилтиадиазолина в этаноле (1), растворе 0.1 M HCl (2) и растворе 0.1 M NaOH (3). Зависимости логарифмов молярных коэффициентов поглощения от длин волн заимствованы из работы [8]. На рис. 6 для спектров 1 и 2 представлена зависимость $\mu_{сум}$ от границ нечеткости $\mu_{сум} = f(\lambda_{L,R}, \lg\varepsilon_{L,R})$. Учитывая, что варьирование $\lambda_{L,R}$ и $\lg \varepsilon_{L,R}$ в интервалах 4-7 нм и 0.09-0.15 соответственно, мало сказывается на величинах $\mu_{сум}$, дальнейшие расчеты проводили, принимая границы нечеткости $\lambda_L = \lambda_R = 5$ нм, $\lg \varepsilon_L = \lg \varepsilon_R = 0.1$. Результаты сравнения спектров приведены в табл. 3.

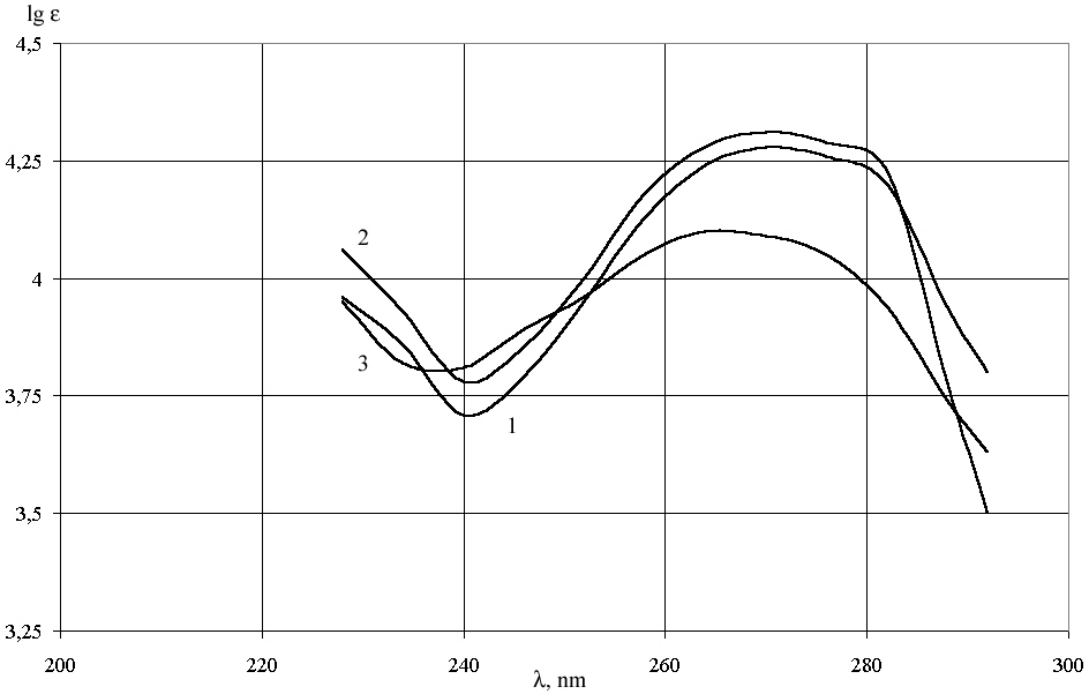


Рис. 5. Спектры поглощения 2-бензолсульфонимидо-5-метилтиадиазолина.

Таблица 3. Критерии сходства спектров 2-бензолсульфонимидо-5-метилтиадиазолина в разных растворителях, $\mu_{сум}$; *d*

		Номер спектра		
		1	2	3
Номер спектра	1	1; 0		
	2	0.80; 0.54	1; 0	
	3	0.34; 0.85	0.33; 0.87	1; 0

Близость значений d друг к другу затрудняет формулировку выводов о принадлежности спектров одному веществу или же о существенном отличии спектров. Авторы работы [7] постулировали, что значения критерия $d > 0.3$ приводят к выводу об неидентичности данных. Принимая разные значения относительных погрешностей откликов, рассчитали величины статистик χ^2 для исследованных систем (табл. 4, 5).

Таблица 4. Значения статистики χ^2 для оценки степени близости спектров бензоилацетона, измеренных в разных растворителях

Относительная погрешность измерения откликов (δ, %)	5	10	15	20
χ^2	8260	2060	920	520

Таблица 5. Значения статистик χ^2 для оценки степеней близости спектров 2-бензолсульфонимидо-5-метилтиадиазолина, измеренных в разных условиях, χ^2 (δ = 5; 10 %)

		Номер спектра		
		1	2	3
Номер спектра	1	0		
	2	7.69; 1.92	0	
	3	17.23; 4.31	17.27; 4.32	0

В случае спектров бензоилацетона, измеренных в воде и в эфире, значения статистик χ^2 превосходят значения 5 %-ной точки распределения χ^2 ($\chi^2_{f=10,\alpha=0.05} = 3.94$), что свидетельствует о значимом различии спектров. Напротив, при сравнении спектров 2-бензолсульфонимидо-5-метилтиадиазолина статистики $\chi^2 < \chi^2_{f=31,\alpha=0.05} = 19.28$, что не противоречит гипотезе об идентичности спектров.

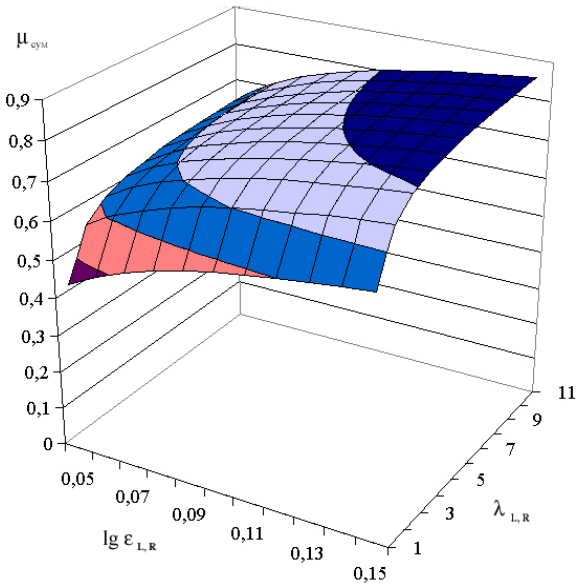


Рис. 6. Зависимость суммарной функции принадлежности от границ нечеткости $\mu_{сум} = f(\lambda_{L,R}, \lg\varepsilon_{L,R})$ для спектров поглощения 2-бензолсульфонимидо-5-метилтиадиазолина в этаноле и в растворе 0.1 М HCl.

Применение критерия сходства, основанного на использовании теории нечетких множеств, в отличие от метода, основанного на расчете Евклидовых расстояний, позволяет выявить отличие спектров 2-бензолсульфонимидо-5-метилтиадиазолина: значение критерия $\mu_{сум} = 0.80$ для спектров 1 и 2 указывает на то, что с высокой вероятностью эти спектры можно считать принадлежащими одному веществу и измеренными в близких условиях. Напротив, спектр 3 или принадлежит другому веществу или условия его измерения существенно отличались от условий измерения спектров 1 и 2.

Накопление данных по оценке сходства позволит, вероятно, сформировать эвристическую базу для выбора критических значений $\mu_{сум}$ и, более того, для формулирования правил идентификации и определении схожести объектов в задачах инструментального качественного анализа. Использованный в данной работе вид функции принадлежности – наиболее простая интерпретация нечеткого числа. Можно ожидать, что применение функций принадлежности типа Гауссовой или Лапласовской увеличит разрешающую способность критерия $\mu_{сум}$.

Авторы выражают благодарность проф. Л.П. Логиновой, доц. Е.А. Решетняк и доц. В.В. Иванову за помощь в работе над рукописью и ряд ценных замечаний, Фонду фундаментальных исследований Харьковского национального университета имени В. Н. Каразина (тема № 17-05), Харьковскому городскому благотворительному фонду Юрия Сапронова и Харьковскому национальному университету имени В. Н. Каразина (тема НИР № 15-15-03) за финансовую поддержку.

Список использованных источников

1. Вершинин В.И., Дерендяев Б.Г., Лебедев К.С. Компьютерная идентификация органических соединений. М.: Академкнига, 2002. – 197 с.
2. Вершинин В.И., Топчий В.А., Медведовская И.И. Критерии совпадения пиков в качественном хроматографическом анализе. Учет воспроизводимости характеристик удерживания // Журн. аналит. химии. – 2001. – Т. 56, № 4. – С. 367-373.
3. Leach A.R., Gillet V.J. An introduction to chemoinformatics. Kluwer Academic publishers, 2003. – 259 p.
4. Дюран Б., Оделл П. Кластерный анализ. М.: Статистика, 1977. – 128 с.
5. Размытые множества / Классификация и кластер. М.: Мир, 1980. – 389 с.
6. Орловский С.А. Проблемы принятия решений при нечеткой исходной информации. М.: Наука. Главная редакция физ.-мат. литературы, 1981. – 208 с.
7. Bandemer H., Otto M. Fuzzy theory in analytical chemistry // Mikrochim. Acta. – 1986. – No. 2. – P. 93-124.
8. Большаков Г.Ф., Ватаго В.С., Агрест Ф.Б. Ультрафиолетовые спектры гетероорганических соединений. Л.: Химия, 1969. – 504 с.

Поступила в редакцию 5 июля 2005 г.

Kharkov University Bulletin. 2005. №669. Chemical Series. Issue 13(36). A. Panteleimonov, Yu. Kholin. Development of fuzzy criteria of identification in qualitative chemical analysis.

The fuzzy sets theory was applied to construct the similarity criteria of objects in the qualitative instrumental analysis. The properties of the criterion proposed were exemplified with the use of the spectroscopic data. Also, the comparison of the new criterion with the existed one based on the estimation of the Euclidean distance between objects was performed.