УДК 519.711

А.М. Вітюк, М.В. Вітюк, В.М. Машін

ЗАСТОСУВАННЯ МЕТОДОЛОГІЇ НЕЧІТКОЇ ЛОГІКИ ДЛЯ ВИРІШЕННЯ ЗАДАЧІ «КІЛЬКІСНИЙ ЗВ'ЯЗОК СТРУКТУРА-ВЛАСТИВІСТЬ»

Застосування нечіткої логіки до модельної вибірки об'єктів для вирішення задачі «зв'язок структура-властивість» дозволило виділити підвибірки активних, неактивних та малоактивних об'єктів та сформувати кластери активних і неактивних об'єктів, максимально рознесені один від одного в багатовимірному просторі описових ознак − тим самим, відповідно, підвищити якість розпізнавання досліджуваної вибірки. Дисперсійний аналіз показав статистичну значимість середніх значень активностей об'єктів кожної з трьох підвибірок. Це дозволило в якості вимірювача міжкласової відстані в просторі описових ознак використати модифікований тренд-вектор. Методом тренд-вектора показано, що в багатовимірному просторі ознак кластер малоактивних об'єктів знаходиться ближче до кластеру активних об'єктів, ніж до кластеру неактивних об'єктів. Тому висловлено припущення, що структурні ознаки малоактивних об'єктів вносять певний внесок у формування досліджуваного відгуку. Інформацію про внесок малоактивних об'єктів можна витягти, користуючись більш поглибленими методами вирішення поставленого завдання для конкретної вибірки реальних об'єктів.

***Ключові слова*:** *кількісний зв'язок структура-властивість (активність), нечітка логіка, описові ознаки, відгук системи, навчальна вибірка, центри систем об'єктів заданої активності у багатовимірному просторі ознак, модифікований тренд-вектор.*

Применение методологии нечеткой логики к модельной выборки объектов для решения задачи «связь структура-свойство» позволило выделить подвыборки активных, неактивных и малоактивных объектов и сформировать кластеры активных и неактивных объектов, максимально разнесенные друг от друга в многомерном пространстве описательных признаков − тем самым, соответственно, повысить качество распознавания исследуемой выборки. Дисперсионный анализ показал статистическую значимость средних значений активностей объектов каждой из трех подвыборок. Это позволило в качестве измерителя межклассового расстояния в пространстве описательных признаков использовать модифицированный тренд-вектор. Показано, что в многомерном признаковом пространстве кластер малоактивных объектов находится ближе к кластеру активных объектов, чем к кластеру неактивных объектов и высказано предположение, что структурные признаки малоактивных

объектов вносят определенный вклад в формирование исследуемого отклика который можно извлечь, пользуясь более углубленными методами решения поставленной задачи для конкретной выборки реальных объектов.

***Ключевые слова*:** *количественная связь «структура-свойство», нечеткая логика, описательные признаки, отклик системы, обучающая выборка, центры систем объектов заданной активности в многомерном признаковом пространстве, модифицированный тренд-вектор.*

The use of fuzzy logic to model non-optimal set of test objects for solving the problem «structure-property relationship» allowed to form subsamples of active, inactive, and low active objects. Its allows to create clusters of active and inactive objects, maximally separated from each other in the multidimensional space of descriptive signs − thus improve the quality of the re-cognition of the investigated objects. Analysis of variance showed a statistical significance of average values of activity of objects in each of the three sub-samples. This allowed to use modified trend-vector as a distance measure in the space of descriptive signs. The method of trend-vector showed that in a multidimensional descriptive space cluster of low active objects is located close to the cluster of active objects than to the cluster of inactive objects. It is suggested that the structural features of low-level objects contributes to the formation of the test response. The information about the contribution of low-level objects can be found by using more in-depth techniques to solve this problem for a specific sample of real objects.

***Keywords*:** *structure-activity relationship, fuzzy logic, descriptive signs, the response of the system, training sample, centers of systems of objects with given activity in a multi-dimensional signs space, modified trend vector*.

В настоящее время разработка методов поиска количественных связей «структура объекта (состав, строение)-его активность (свойство)», то есть, решение задачи QSA(P)R («Quantitative Structure-Activity (Property) Relationship», *англ.*) выделилась в отдельное научное направление, решение которого имеет важное значение для современных технологий. Сначала задача QSAR была сформулирована химиками. Решение этой задачи позволяет без синтеза молекул некоторого химического ряда прогнозировать их отклик (например, биологическую активность) по заданным структурным характеристикам и проводить оценку свойств новых соединений вне эксперимента[1]. Решение QSAR задачи имеет все специфические черты теории распознавания образов и поэтому подходы, которые используются для решения этой задачи, применяются в решении не только химико-биологических задач, но и многих технических задач [2-3].

QSAR-моделирование позволяет проводить обобщение разноплановых экспериментальных данных, накапливаемых в компьютерных структурных базах данных и получать ответы на вопросы о том, как следует оптимизировать решение поставленной задачи .

Успешное применение теории распознавания образов требует корректного построения обучающей выборки − то есть, набора *i* объектов, для каждого из которых, исходя из значения его отклика (A_i), известно *a priori* к которому из нескольких классов он принадлежит,

В решении QSAR задачи обучающая выборка обычно состоит из двух подвыборок: активных объектов (такую выборку из А-элементов, для которых рассматриваемый отклик $A_i>A_{гр}$, будем называть А-выборкой) и подвыборки неактивных объектов (Н-выборка, образованная Н-элементами для которых $A_i<A_{гр}$). Понятие «разделитель» $A_{гр}$ формулируется исследователем на основе эмпирически накопленных данных и в процессе решения конкретной практической задачи QSAR-задачи понятие «разделителя» и его численное выражение $A_{гр}$ могут быть изменены.

Для описания структуры объекта обычно используются два вида характеристик: локальные, показывающие наличие определенного фрагмента в составе объекта и интегральные, характеризующие весь объект в целом. Для описания исследуемой выборки объектов применяется многомерное признаковое пространство, при этом используются различные шкалы измерений описательных признаков S_i, и отклика A_i, в том числе и бинарная − присутствие/отсутствие данного признака S_{ij} (индекс j − нумерует признаки) кодируется, соответственно, единицей или нулем.

Известные методы преобразования многомерного признакового пространства в двумерное, и его классификации [4-6] основаны на разделении исследуемой выборки объектов на А- и Н-подвыборки и требуют корректного подхода к заданию «разделителя» $A_{гр}$.

Фундаментальный закон естествознания − принцип целостности − утверждает, что в биосфере любой объект является единой коммуникативной системой, в которой «все связано со всем» [7]. Отсюда следует, что парные корреляционные связи между описательными признаками S_j *i*-того объекта уменьшают информативность взаимно коррелирующих признаков и затрудняют использование классических регрессионных методов нахождения функциональной связи «структура-свойство».

Привлекательной особенностью методологии нечеткой логики [8-10] является определение принадлежности объекта к заданному классу на основе нечеткого описания признаков каждого объекта с использованием лингвистических переменных, отражающих взаимоотношение некоторого признака с остальными признаками данного объекта. При введении лингвистических переменных исследователем используется *а приорная* информация. Методология нечеткой логики дает возможность «обратной связи» между объектом и субъектом путем эволюционного «переобучения» (переформатирования) А- и Н-подвыборок по результатам предыдущего этапа решения данной задачи, при неизменности коренного алгоритма решения задачи. В частности, такое «переобучение» исходной выборки может состоять из выделения из неё подвыборок с заданным интервалом величины отклика. В методологии нечеткой логики возможна

ситуация, когда объект может принадлежать к по своим описательным структурным признакам двум классам активности, т.е. множества значений классов пересекаются. Из этого следует, что при распознавании образов следует рассматривать также промежуточный класс малоактивных объектов, которые согласно принципу целостности, могут нести некоторую информацию о влиянии структуры на данный отклик.

Целью настоящей работы является использовать методологию нечеткой логики для образования кластеров объектов, центры систем которых максимально разнесены друг от друга в многомерном пространстве описательных признаков, то есть, повысить качество разделения объектов.

В качестве меры разделения предлагается использовать модифицированный метод тренд-вектора. Тренд-вектор первоначально был предложен для решения задачи QSAR [4-5], поскольку цифровой материал в химии является отображением различных законов, которые описывают свойства соединений, структурную информацию молекул и т.д. В дальнейшем тренд-вектор нашел применение для решения разнообразных задач, в том числе и прикладных технических. Эвристичность метода тренд-вектора следует из того, что в нем также используется основная идея теории распознавания образов – разнесение объектов на два класса А и Н тренд-вектор является аналогом известного вектора дипольного момента $\boldsymbol{p}$ в электростатике, Модуль вектора $\boldsymbol{p}$ равен произведению расстояния между центрами двух систем электрических зарядов (одинаковых по модулю, но разноименных по знаку) умноженному на величину заряда системы.

Аналогично этому, будем рассматривать модифицированный тренд-вектор, соединяющий в многомерном пространстве описательных признаков центры систем L- и M- подвыборок, а за численное значение тренд-вектора (Т) примем топологическое расстояние (Δr) между центрами L- и M- систем, умноженное на разность средних величин активностей объектов этих систем

$$\mathrm{T} = \Delta r\left(A_{cp}^{L} - A_{cp}^{M}\right). \qquad (1)$$

Методология нечеткой логики применена к модельной «неоптимально» спланированной выборки, состоящей из 24 гипотетических объектов, описанных семью «структурными» признаками. В качестве отклика рассматривалась некая «активность» А объектов (табл.1). В табл. 1 исследуемые образцы расположены по убыванию их активности, следовательно, номер образца можно рассматривать как убывающий ранг его активности (абстрагируясь от понятия связанных рангов).

Поскольку для описания структуры (состава) S реальных объектов используются различные шкалы измерений с широким интервалом численных значений, то описательные признаки *S* объектов «модельной» выборки находились в единичном интервале [0,1], при этом использовалось преобразование

$$S^o = (S_{max} - S_i)/(S_{max} - S_{min}), \quad (2)$$

где S_{max} и S_{min} – соответственно максимальное и минимальное значения *j*-того описательного признака *i*-го объекта.

Таблица 1

Описательные признаки и активность модельных объектов

Номер п/п	Описательные признаки							Активность А
	I	II	III	IV	V	VI	VII	
1	1	0,6	0,5	0,4	0,3	0,4	0,5	26
2	0,6	0,5	0,4	0,6	0,5	0,5	0,3	25
3	0,8	0,4	0,6	0,2	0,3	0,3	0,6	24
4	0,4	0,5	0,3	0,5	0,5	0,5	0,4	24
5	0,6	0,6	0,4	0,4	0,8	0,6	0,5	23
6	0,3	0,3	0,6	0,4	0,3	0,4	0,4	22
7	0,5	0,3	0,3	0,4	0,4	0,2	0,3	22
8	0,4	0,5	0,5	0,6	0,3	0,8	0,3	22
9	0,3	0,5	0,4	0,7	0,6	0,6	0,8	21
10	0,4	0,7	0,7	0,7	0,7	0,7	0,6	20
11	0,5	0,7	0,5	0,7	0,6	0,5	0,6	20
12	0,6	0,7	0,7	0,8	0,5	0,6	0,6	18
13	0,4	0,5	0,4	0,4	0,5	0,6	0,4	18
14	0,3	0,6	0,3	0,7	0,6	0,6	0,5	17
15	0,5	0,3	0,7	0,6	0,4	0,6	0,7	17
16	0,5	0,4	0,5	0,5	0,8	0,7	0,3	17
17	0,7	0,6	0,7	0,5	0,3	0,7	0,3	16
18	0,6	0,5	0,4	0,7	0,5	0,4	0,4	12
19	0,8	0,6	0,3	0,6	0,7	0,5	0,3	10
20	0,5	0,5	0,4	0,4	0,4	0,4	0,4	8
21	0,2	0,1	0,2	0,3	0,2	0,5	0,3	6
22	0,1	0,4	0,3	0,5	0,7	0,4	0,5	6
23	0,4	0,2	0,6	0,4	0,2	0,2	0,2	4
24	0,5	0,3	0,2	0,1	0,3	0,2	0,4	4

Таблица 2

Ранговая корреляция (ρ) по Спирмену между описательными признаками I-VII исходной выборки и ранговая корреляция ($\rho_{Акт}$) их с активностью

Описательные признаки								$\rho_{Акт}$
	I	II	III	IV	V	VI	VII	
I	1,000	0,374	0,280	-0,029	-0,015	-0,088	-0,020	0,262
II	0,374	1,000	0,247	**0,615**	**0,517**	**0,496**	**0,371**	0,300
III	0,280	0,247	1,000	0,263	-0,135	0,342	0,268	0,223
IV	-0,029	**0,615**	0,263	1,000	**0,540**	**0,543**	**0,345**	0,069
V	-0,015	**0,517**	-0,135	**0,540**	1,000	**0,412**	0,330	0,086
VI	-0,088	**0,496**	0,342	**0,543**	**0,412**	1,000	0,140	0,148
VII	-0,020	**0,371**	0,268	**0,345**	0,330	0,140	1,000	0,260

О «неоптимальности» сконструированной базы исходных данных свидетельствует табл. 2 (выделены значимые коэффициенты корреляции), которая показывает отсутствие описательных признаков индивидуально значимо коррелирующих с активностью объектов (для N = 24 на уровне α = 0,05 $\rho_{крит}$ = 0,344 [11]) и наличие взаимно коррелирующих описательных признаков.

На первом этапе решения поставленной задачи модельную выборку подвергли многомерному кластерному анализу.

Дендрограмма (рис. 1) на основе матрицы сходства пар объектов изображает взаимные связи между описательными признаками объектов исследуемой выборки. Оценивалось евклидово расстояние между парами исследуемых объектов в семимерном описательном пространстве.

Из рис.1 видно, что в исследуемой выборке имеются объекты, которые в семимерном признаковом пространстве «топологически» близки друг к другу, но характеризуются резко отличающейся номерами (активностью). Например, топологические расстояния между объектами № 2 и № 18, № 4 и № 13 менее 0,2, тогда, как топологическое расстояние между наиболее активными объектами № 1 и № 3 составляет 3,8.

Методом кластер-анализа провели разделение исследуемых объектов на два класса, а затем и на три класса. При методе «k-средних» каждый объект относится к тому кластеру, к центру которого он ближе всего для того чтобы минимизировать суммарное квадратичное отклонение точек кластеров от центра данного кластера.

При делении на два кластера было получено, что кластер № I состоит из 16 объектов (№ 2, 4, 5, 8-19, 22 со средней активностью $A_{ср}$ = 17,9), а кластер № 2 состоит из 8 объектов (№ 1, 3, 6, 7, 20, 21,23, 24; $A_{ср}$ = 14,5).

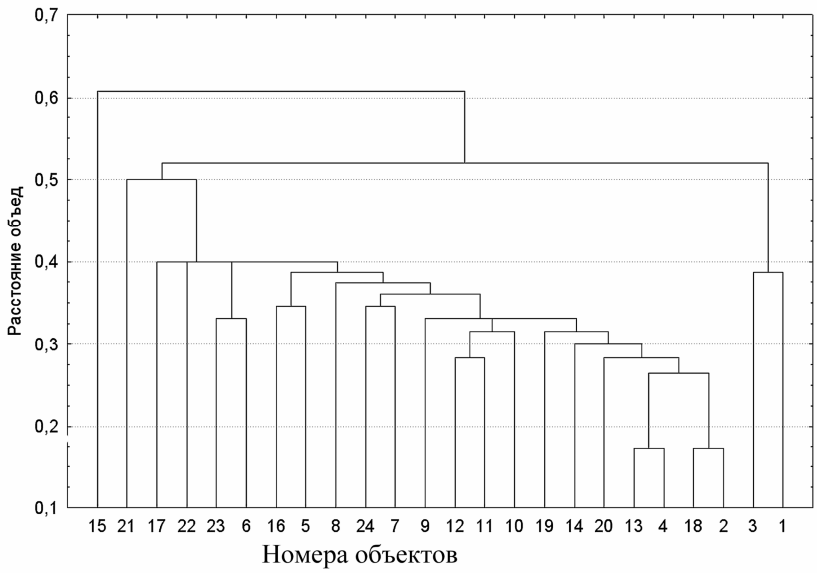

Рис.1. Иерархическая классификация исследуемых 24 объектов

При делении на три кластера были получены: кластер № 1 содержит 6 объектов (№ 3, 6, 7, 21, 23, 24; $A_{ср}$=13,7), кластер № 2 содержит 11 объектов (№ 1, 2, 4, 5, 8, 13.16-20; $A_{ср}$= 18,3), кластер № 3 содержит 17 объектов (№ 9-12, 14, 15, 22; $A_{ср}$= 18).

Однако, дисперсионный анализ для проверки нескольких средних по одному признаку (активности) показал, что различия в средних активностях выделенных кластер-анализом подвыборок статистически не значимы и принимается нулевая гипотеза, о том, что средние активности объектов подвыборок равны. Следовательно, применение кластер-анализа для разбиения исследуемой выборки на подвыборки активных и неактивных объектов не приводят к статистически значимым результатам.

Методология нечеткой логики в отличие от метрических алгоритмов распознавания образов (например, метод k-ближайших соседей в кластер-анализе) использует описание структуры объекта, при котором каждый структурный признак связан с остальными признаками данного объекта логическими переменными. Такое описание структуры объекта обладает «синергизмом» совместного использования локальных и интегральных описательных признаков объекта, что повышает качество описания структуры объекта в целом. Использование методологии нечеткой логики, позволяет *a priori* допустить, что в модельной выборке присутствуют объекты, которые по величине своей активности не могут быть «четко» приписаны к определенному классу. Действительно, методология нечёткой логики позволила сформировать три подвыборки «модельных» объектов с пересекающейся активностью (табл. 4): активная (А-выборка, 17<A_i, N_A =13), малоактивная (М-выборка, 10<A_i<24, N_M =10) и неактивная (Н-выборка, (A_i<17, N_H = 8).

Таблица 4

Разнесение объектов методологией нечеткой логики на три подвыборки

Подвыборка	Номера объектов, включенных в подвыборку	Средняя активность элементов подвыборки
А-выборка	№ 1-13	21,92
М-выборка	№ 5, 6; № 11-18	18,00
Н-вывборка	№ 17-24	7,25
«Пересекающиеся» объекты: № 5,6 отнесены и к А- и к М-выборкам; Объекты № 17-18 отнесены и к М- и к Н-выборкам		

Дисперсионный анализ полученных трех подвыборок показал значение критериальной статистики $F = 74{,}11$.

Полученная величина является значимой [11]. Нулевую гипотезу, состоящую в том, что активность объектов совпадает во всех трех группах, следует отбросить.

Статистическая значимость средних значений активностей ($A_{ср}$) объектов в сформированных подвыборках позволяет при помощи модифицированного Т-вектора определить разделяемость А-, Н- и М- подвыборок. Значения Т-векторов получены на основании данных, представленных в таблице 5.

Таблица 5

Средние значения описательных признаков и топологические расстояния Δr между подвыборками. Разности средних активностей объектов подвыборок и соответствующие Т-вектора

Средние значения описательных признаков по выборкам								Δr	Средняя активность выборок, значения Т-векторов
Выборка	I	II	III	IV	V	VI	VII		
А	0,523	0,523	0,485	0,523	0,485	0,515	0,485		21,92
М	0,500	0,520	0,520	0,570	0,530	0,570	0,470		18,00
Н	0,475	0,400	0,388	0,438	0,413	0,413	0,350		8,25
А-М	0,023	0,003	-0,035	-0,047	-0,045	-0,055	0,015	0,096	$\mathbf{T_{АМ}}$ = (21,92-18,00) ·0,096= **0,376**
А-Н	0,048	0,123	0,097	0,085	0,072	0,103	0,135	0,261	$\mathbf{T_{АН}}$ = =(21,92-8,25) ·0,261=**3,568**
М-Н	0,025	0,120	0,132	0,132	0,117	0,157	0,120	0,321	$\mathbf{T_{МН}}$ = (18,00-8,25) ·0.321=**3,130**

Значения Т-векторов свидетельствуют, что в многомерном пространстве «структура объекта – отклик» максимально разделены активная и неактивная подвыборки исследуемых объектов (T_{AH} = 3,568). Обращает на себя внимание тот факт, что в многомерном пространстве «структура-отклик») малоактивная выборка находится ближе к активной, чем к неактивной (T_{AM} = 0,376<T_{MH} = 3,130).

На основании изложенного можно предположить, что структурные характеристики малоактивной подвыборки несут некоторую информацию, которая способствует проявлению данного отклика, для выявления этой информации требуется применение более глубоких методов анализа при решении конкретных прикладных задач.

СПИСОК ЛІТЕРАТУРИ

1. *Стьюпер Э., Брюггер У., Джурс П. Машинный анализ связи химической структуры и биологической активности: Пер. с англ. − М.: Мир, 1982. − 240 с.*
2. *Журавлев Ю.И., Гуревич И.Б. Распознавание образов и анализ изображений / В сб. Искусственный интеллект. Справочник. −Т. 2 / Под ред. Д.А. Поспелова. − М.: Наука, 1990. − С.149-190.*
3. *Немчук О.О., Вітюк М.В., Матоліков Д.П. Метод аналізу параметрів систем автоматизації приводів перевантажувачів на основі мір схожості // Теорія і практика будівництва. − 2012. − № 7. − С.33-39.*
4. *Carhart R.E. Atomic Pairs as Molecular Features in Structure-Activity Studies: Definitions and Applications / R.E. Carhart, D.H. Smith, R. Venkataraghavan // Journal of Chemical Information and Computer Sciences. − 1985. – V. 25(2). – P.64-73.*
5. *Sheridan R.P.I. Extending the trend vector: the trend matrix and sample-based partial least squares / R.P.I. Sheridan, R.B.Nachbar, B.I.Bush // Journal Computed Aided Molecular Desay. – 1994. – V. 8(3). – P. 64-73.*
6. *Кузьмин В.Е. Механистические модели в хемометрике для анализа многомерных исследовательских данных / В.Е. Кузьмин, Н.В. Витюк // Журнал аналитической химии. − 1994. – Т. 49. – № 2. – С. 165-172.*
7. *Коммонер Б. Замыкающийся круг. – М.: Гидрометеоиздат, 1974. – 280 с.*
8. *Штовба С.Д. Проектирование нечетких систем средствами MATLAB. – М.: Горячая линия – Телеком, 2007. – 288 с.*

9. *Деветьяров Д.А. Эволюционное построение алфавита дескрипторов, сформированных на основе аппарата нечеткой логики, в задаче «структура-свойство» // Системы управления и информационные технологии. – 2010. – № 1.1 (39). – С. 131-134.*
10. *Козловский В.А., Максимова А.Ю. Нечеткая система распознавания образов для решения задачи классификации жидких нефтепродуктов // Наукові праці ДонНТУ. Серія «Інформатика, кібернетика та обчислювальна техніка». – Вип. 13(185). – 2011.*
11. *Поллард Дж. Справочник по вычислительным методам статистики / Пер. с англ. В.С. Занадворова; Под ред. и с предисл. Е.М. Четыркина. – М.: Финансы и статистика, 1982. – 344 с.*

Стаття надійшла до редакції 12.12.2016

Рецензенти:

доктор технічних наук, професcор, завідувач кафедри «Інформаційні технології» Одеського національного морського університету **В.В. Вичужанін**

кандидат технічних наук, ст. викладач кафедри «Мережі та системи поштового зв'язку» Одеської національної академії зв'язку (ОНАЗ) ім. О.С. Попова **В.Ю. Кумиш**