РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ
А В Т О М Е Т Р И Я
№ 5 2001

УДК 519.816

В. И. Красинский
(Новосибирск)

НЕЧЕТКАЯ КЛАССИФИКАЦИЯ ОБЪЕКТОВ МАЛОЙ ЧИСЛОВОЙ ВЫБОРКИ

Статистические методы классификации не применимы к короткому вариационному ряду, построенному по числовому признаку биологических объектов (тепловому индексу 26 видов растений). Применена гипотеза λ-компактности Н. Г. Загоруйко для разбиения множества объектов на классы. Далее на основе теории нечетких множеств осуществляется вычисление степеней принадлежности объектов к классам согласно гипотезе автора о переменной неразличимости объектов. Приведены исходные данные и алгоритм вычисления по ним функций принадлежности объектов к нечетким множествам (классам) относительно трех значений лингвистической переменной требовательности растений к теплу. Дан пример получения нечетких продукционных правил для построения экспертных систем. Описанные результаты исследования соответствуют представлениям ведущих специалистов предметной области. Показана пригодность предложенной методики нечеткой классификации малого числа объектов для тех многочисленных случаев обработки экспериментальных данных, в которых числовой признак измерен в шкале отношений.

1. **Постановка задачи.** В данной работе анализируется вариационный ряд биологических объектов (видов растений), построенный по количественному физиологическому показателю теплового индекса (ТИ), исходные данные взяты из [1]. Этот показатель характеризует степень требовательности растений к теплу (ТРТ), соответственно интенсивность метаболизма, и классификация видов растений по этому признаку признается биологами весьма актуальной для решения прикладных задач интродукции и селекции растений.

Вариационный ряд из 26 объектов по возрастанию признака ТИ представлен в таблице. Для восьми из этих объектов имеются, кроме того, достоверные экспертные оценки по трехбалльной шкале степени ТРТ. Эти маркерные объекты помечены сокращенным именем своего класса в колонке «Эксперт». Представленный вариационный ряд биологи весьма произвольно разделяют на три класса по значениям ТИ, следя лишь за тем, чтобы маркерные объекты попали в соответствующие классы по степени ТРТ.

В таблице приведены ботанические названия видов (объектов) и производные от значений ТИ показатели, которые будут объяснены далее. На рис. 1 приведена гистограмма объектов по значениям ТИ. Видно, что об аппроксимации этой двумодальной гистограммы каким-либо известным законом распределения вероятностей не может быть и речи. Количество объек-

№ п/п	Виды растений	Эксперт	ТИ	*D*	λ
	холодовыносливые (х)				
1	Sedum pachyphy		3,20	0,00	0,00
2	Brassica oleracea	х	3,76	0,56	6,22
3	Pulmonaria mollis		3,85	0,09	1,29
4	Bergenia crassifolia	х	3,92	0,07	0,78
5	Corydalis bracteata		4,08	0,16	2,29
6	Pinus sylvestris		4,17	0,09	1,50
7	Helianthus tuberosus		4,23	0,06	3,00
8	Vicia faba	х	4,25	0,02	0,33
9	Beta vulgaris	х	4,34	0,09	4,50
10	Aloe arborescens		4,45	0,11	5,50
11	Euphorbia secucunien		4,47	0,02	0,18
	теплолюбивые (т)	**4,53 – граница х–т**			
12	Daucus carota		4,72	0,25	12,50
13	Secale cereale		4,78	0,06	3,00
14	Angelica sylvestris		4,80	0,02	0,67
15	Betula pendula		4,83	0,03	1,50
16	Artemisia dracunculus		4,89	0,06	2,00
17	Heracleum dissectum		5,10	0,21	3,50
18	Zea mays	т	5,26	0,16	2,29
19	Pteridium aquilinum		5,33	0,07	1,00
20	Tilia cordata		5,40	0,07	1,75
21	Phaseolus vulgaris	т	5,44	0,04	0,57
	весьма теплолюбивые (ВТ)	**5,65 – граница т–ВТ**			
22	Philodendron selloum		6,03	0,59	14,75
23	Cucumis sativus	ВТ	6,10	0,07	0,32
24	Diffenbachia seguine		6,32	0,22	3,14
25	Monstera deliciosa		6,50	0,18	0,82
26	Gossypium hirsutum	ВТ	6,86	0,36	2,40

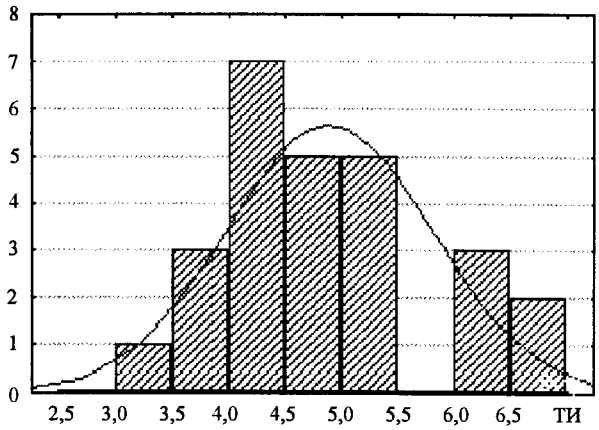

Рис. 1. Гистограмма по значениям ТИ объектов таблицы: плавная линия – функция плотности нормального закона распределения вероятностей

тов мало, поэтому нет оснований применять для анализа (классификации) какой-либо статистический метод. Проведем анализ показателя ТИ при помощи некоторых эвристических процедур, применяемых при распознавании образов [2], привлекая теорию нечетких множеств [3], теорию возможностей [4, 5] и один из методов обработки парных сравнений объектов [6].

Целью работы являются, во-первых, формализация процесса классификации объектов по значениям признака ТИ, во-вторых, определение степеней одновременных принадлежностей объектов к смежным классам. Решение первой задачи осложнено малым объемом исходных данных (короткий вариационный ряд). Решение второй задачи важно с методической и практической точек зрения, поскольку четкая классификация объектов является значительным огрублением предметной области. Таким образом, цель исследования – решение типовой задачи обработки результатов эксперимента в затрудненных условиях.

Для анализа в разд. 4 нечеткой принадлежности объектов к классам необходимо определить тип шкалы, в которой измерены значения признака. Согласно методике предварительной обработки результатов физиологического эксперимента, инвариантный к большинству природных условий безразмерный индекс ТИ есть отношение количеств поглощенного растением кислорода (интенсивность метаболизма) при двух фиксированных температурах. Такая интерпретация результатов эксперимента соответствует измерению признака ТИ в шкале отношений [7]. Эта шкала характеризуется допустимым преобразованием подобия $\varphi(y) = ky$. В ней измеряются, например, масса тела, цена товара по отношению к эталону.

2. **Разбиение объектов на четкие классы.** В условиях отмеченной выше неприменимости статистических методов классификации объектов воспользуемся обоснованной в [2] гипотезой λ-компактности для разбиения объектов на классы. В рассматриваемом одномерном случае применение гипотезы для числового признака несложно.

Согласно этому методу объекты отождествляются с точками на оси значений некоторой меры объектов, а критерием, дискриминирующим классы объектов, считается максимальное λ-расстояние между точками (объектами). Это λ-расстояние есть отношение расстояния между точками, соответ-

ствующими положению объектов на числовой оси меры, к минимальному из двух соседних расстояний. Расстояния ΔТИ между объектами в таблице занесены в колонку «D».

Поясним на примере. Рассмотрим в таблице объекты 3–6. Расстояние $D_{45} = 0{,}16$, а соседние $D_{34} = 0{,}07$ и $D_{56} = 0{,}09$. Вычислим отношения: $D_{45}/D_{34} = 2{,}29$, $D_{45}/D_{56} = 1{,}78$. Максимальное из этих отношений 2,29 принимается за λ-расстояние между объектами 4 и 5. Отношение расстояний подчеркивает (контрастирует) неоднородность структуры расстояний между объектами. Итак, λ-расстояние между двумя объектами есть максимальное из двух соседних относительных расстояний. В таблице это расстояние занесено в колонку «λ» для каждой пары объектов.

По критерию максимальных λ-расстояний в работе [2] предлагается и обосновывается разделение объектов на классы. Число классов выбирается исходя из сущности решаемой задачи. В нашей задаче требуется разделить объекты на три класса. Максимальное значение λ в таблице оказалось между объектами 21 и 22, следующее по величине – между объектами 11 и 12. Таким образом, разделение на три класса произведено.

По существу биологической задачи виды растений 1–11 относятся к холодовыносливым, виды 12–21 относятся к теплолюбивым, виды 22–26 относятся к весьма теплолюбивым. Все маркерные объекты при этом попали в свои классы. Это является хорошим подтверждением, с одной стороны, физиологической методики получения показателя ТИ, а с другой – применимости гипотезы λ-компактности для классификации реальных объектов.

Можно было бы сравнить разбиение на три класса методом максимальных λ-расстояний с разбиением другими методами, но такое исследование выходит за рамки данной работы, тем более что полученное разбиение вполне удовлетворяет исследователей-прикладников (биологов). Кроме того, изложенный далее способ определения нечеткой принадлежности объектов к классам не зависит от самих границ классов.

3. **Уточнение четких границ между классами.** В разд. 2 описан алгоритм разбиения объектов на классы, но не определены точные границы между классами, или пороги дискриминации. Установление порогов важно при решении задачи классификации новых объектов по значению признака ТИ, считая таблицу обучающей выборкой. Также очень важно оценить степень уверенности в классификации «приграничных» объектов. Например, маркерный объект 21 находится непосредственно вблизи границы своего класса и вследствие ошибки в измерениях признака вполне может оказаться в соседнем классе.

Границы между классами уточним эвристически, на основе тех же расстояний между объектами (ΔТИ), по которым вычислялись λ-расстояния, а именно установим точные границы между классами в точках, разделяющих расстояние между пограничными объектами пропорционально величинам соседних расстояний (ребер). Полученные таким образом две границы между тремя классами 4,53 и 5,65 записаны в колонке «ТИ» таблицы в строках с названиями классов «теплолюбивые» и «весьма теплолюбивые». Следовательно, значение ТИ = 4,53 является четкой границей между классами «холодовыносливые» и «теплолюбивые», а значение ТИ = 5,65 есть четкая граница между классами «теплолюбивые» и «весьма теплолюбивые». На рис. 2 эти границы обозначены точками S и G.

4. **Определение степеней принадлежностей объектов к классам.** Очевидно, что четкое разделение объектов на классы по измеренным значениям

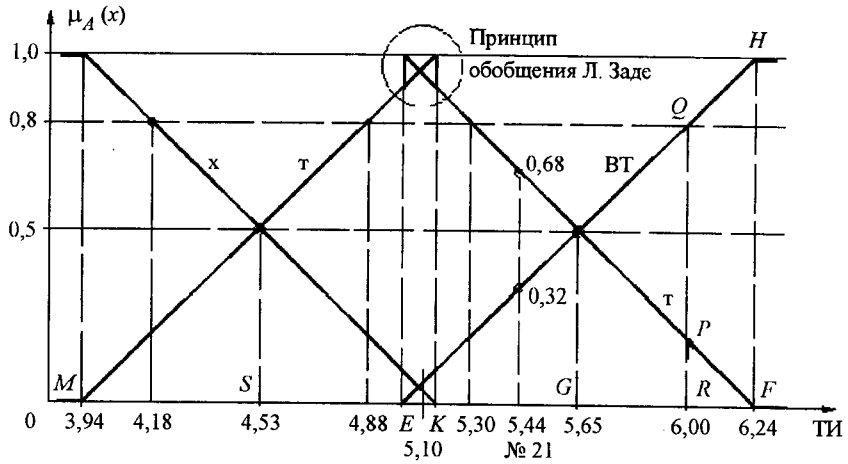

Рис. 2. Графики функции принадлежности нечетких множеств степеней ТРТ

некоторого признака является значительным упрощением предметной области. При решении практических задач желательна более реалистичная оценка степени принадлежности объектов к классам.

Подходящей моделью для учета относительной принадлежности объектов к классам являются предложенные Л. Заде [3] нечеткие множества (НМ). Теория НМ позволяет в формализованном виде манипулировать так называемыми лингвистическими переменными для адекватного отражения реальных явлений. В решаемой задаче такой лингвистической переменной является понятие требовательности растений к теплу с порядковыми значениями (термами) этой переменной «холодовыносливые», «теплолюбивые», «весьма теплолюбивые», что отражено в таблице.

Каждый терм лингвистической переменной ТРТ отождествляется со своим НМ. Степень принадлежности объекта к нечеткому множеству оценивается значением функции принадлежности (ФП), обозначаемой обычно $\mu_A(x)$ и принимающей значения в интервале [0, 1]. В нашей задаче значения трех ФП каждого объекта характеризуют степени принадлежности этого объекта к соответствующим классам по значениям лингвистической переменной ТРТ, т. е. фактически решается задача нечеткого дискриминантного анализа.

Одной из основных проблем применения НМ для оценки реальных ситуаций является как раз вычисление значений ФП объектов. Часто их задают экспертно, называя субъективными вероятностями. Такой способ вызывает наибольшую критику в адрес теории нечетких множеств. Вычислим все значения ФП, не прибегая к экспертным оценкам для каждого объекта, а основываясь на исходных данных таблицы. Из этих данных надо извлечь критерий нечеткости для вычисления $\mu_A(x)$.

Критерий нечеткости границ между классами, или, что то же самое, нечеткой принадлежности объектов к классам, будем искать в рамках теории возможностей. Категориями «необходимость» и «возможность» оперировал еще Аристотель. Его постулат, «если некоторое событие необходимо, то противоположное событие невозможно», и в современной математической теории возможностей, предложенной Л. Заде [4], является основополагающим. Теория ошибок является частным случаем теории возможностей, если степени возможности некоторого события принимают лишь значения 0 или 1.

Переменная возможность события отождествляется с некоторым нечетким множеством. В нашей задаче возможностью события A назовем возможность различимости объектов по значению признака ТИ. Величину ΔТИ в колонке «D» таблицы примем в качестве критерия для вычисления ФП всех объектов к нечетким множествам – термам (х), (т), (ВТ) лингвистической переменной ТРТ. Из таблицы видно, что $D_{\min} = 0{,}02$, $D_{\max} = 0{,}59$. Поэтому можно сформулировать два тезиса по принципу Аристотеля:

– если расстояние ΔТИ между объектами менее $D_{\min}$, то эти объекты не различимы;

– если расстояние ΔТИ между объектами более $D_{\max}$, то эти объекты различимы.

Оба эти тезиса верны для четкого случая. Требуется их совместить и получить переменную степень различимости объектов. В теории ошибок обычно применяют осторожную стратегию, рассчитывая на наихудший случай. Для нашего анализа такой подход можно использовать, приняв в качестве меры возможной неразличимости объектов именно максимальную разность между двумя соседними значениями вариационного ряда. Примем величину $D_{\max} = 0{,}59$ как максимальную меру неразличимости объектов для согласования с принципом необходимости события по Аристотелю (неразличимость объектов с необходимостью заключена в этом пределе), но возможность неразличимости объектов – переменная величина по Л. Заде.

Итак, семантику нечеткой принадлежности объектов к классам, или *гипотезу неразличимости*, формулируем следующим образом: «если расстояние между объектами менее $D_{\max}$, то каждый из них принадлежит с некоторыми степенями к разным классам».

Далее, примем вычисленные выше границы между классами 4,53 и 5,65 за значения признака ТИ условных пограничных объектов и распространим от этих условных объектов величину $D_{\max}$ в обе стороны. Иными словами, середину отрезка $2D_{\max}$ «приложим» к границам классов (точкам S и G на рис. 2) и будем считать возможность $\mu_A(x)$ неразличимости реальных объектов с условными объектами (границами классов) переменной величиной (линейной функцией). Значение этой функции на границах классов примем за 0,5 (равная принадлежность объекта к соседним классам).

В результате приведенных рассуждений получаются точки на горизонталях $\mu_A(x) = 0$ и $\mu_A(x) = 1$ для построения линейных функций принадлежности объектов к трем НМ (термам лингвистической переменной ТРТ). Точности линейной интерполяции ФП вполне достаточно для большинства практических задач.

Эти точки на рис. 2 следующие:

$$M\colon\ 3{,}94 = 4{,}53 - D_{\max}, \qquad E\colon\ 5{,}06 = 5{,}65 - D_{\max},$$

$$K\colon\ 5{,}12 = 4{,}53 + D_{\max}, \qquad F\colon\ 6{,}24 = 5{,}65 + D_{\max}.$$

Графики трех ФП изображены на рис. 2.

Предложенный способ вычисления ФП объектов является модернизацией процедуры построения ФП на основе качественного парного сравнения степеней принадлежности по методике Т. Саати [6]. По этой методике результатом опроса эксперта(ов) является матрица $\|m_{ij}\|$, $i, j = 1, \ldots, n$, где n – число точек (объектов), в которых сравниваются значения ФП. Числа m_{ij} показывают, во сколько раз, по мнению эксперта, степень свойства у объекта i больше,

чем у объекта j. По определению принимается $m_{ii} = 1$ и $m_{ij} = 1/m_{ji}$. Числа для степеней сравнения применяются не произвольные, а из обоснованной психологами универсальной шкалы. Шкала оперирует степенями сравнения свойства исследуемого объекта со свойством некоторого эталонного (среднего) объекта и учитывает эти степени баллами от 1 до 9.

Эвристическая гипотеза метода парных сравнений состоит в том, что отношение баллов свойства объектов, полученных по универсальной шкале, равно отношению значений ФП этих объектов к нечеткому множеству по рассматриваемому свойству, т. е. $m_{ij} = \mu_A(x_i)/\mu_A(x_j)$.

В нашем случае числового признака, измеренного в шкале отношений, в качестве отношений баллов универсальной шкалы примем отношение разностей значений признака объектов. Поясним с помощью рис. 2. Рассмотрим подобные треугольники *ERQ* и *EFH*. Точку *R* будем считать значением ТИ произвольного объекта x. Соответственно отрезок $RQ = \mu_A(x)$, а отрезок $FH = 1$. Отмеченная координата границы между классами $G = 5{,}65$ привязывает треугольник *EFH* к числовой оси признака: $E = G - D_{\max}$. Из пропорции $RQ/FH = ER/EF$, подставляя значение признака, получаем

$$\mu_A(x) = (x - (G - D_{\max}))/2D_{\max}. \qquad (1)$$

В эту формулу входит один параметр – координата G (или S) четкой границы классов, вычисленная ранее.

Вариант для убывающей ФП соответствует принадлежности объекта к дополнительному НМ:

$$\mu_{\neg A}(x) = 1 - (x - (G - D_{\max}))/2D_{\max}. \qquad (2)$$

Итак, в случае принятия гипотезы о величине $D_{\max}$ как максимальной меры переменной линейной неразличимости объектов получаются простые формулы (1), (2) для вычисления ФП объектов к НМ по значению числового признака.

Отметим характерные особенности полученных ФП:

1. Пограничные условные объекты со значениями ТИ = 4,53 и 5,65 имеют равные значения ФП $\mu_A(x) = \mu_B(x) = 0{,}5$ принадлежности к соседним классам, что вполне логично и определено методом построения ФП.

2. Надежную принадлежность при $\mu_A(x) = 1$ к одному классу имеют лишь объекты 1–4 и объекты 24–26. Большинство объектов имеют принадлежность к двум классам. Так, расположенный вблизи границы маркерный объект 21 (Phasoleus vulgaris – фасоль) принадлежит к классу «теплолюбивые» с уверенностью 0,68 и к классу «весьма теплолюбивые» с уверенностью 0,32. Объект 17 со значением ТИ = 5,10 оказался формально принадлежащим сразу к трем классам, хотя превалирует принадлежность к классу «теплолюбивые».

3. В теории НМ доказана так называемая *теорема представления*, согласно которой НМ является совокупностью четких множеств α-уровня: $A_\alpha = \{x \mid \mu_A(x) \geq \alpha\}$. Это множества таких объектов, у которых значение ФП к НМ не менее α. Уровни α есть изолинии степени уверенности в том событии, семантика которого соответствует НМ. Такое представление очень удобно для формулировки *нечетких продукционных правил* (нечеткой импликации), применяемых при построении экспертных систем.

Зададимся вопросом, какие объекты находятся в своих классах с уверенностью более 0,8? Подобные проблемы нечеткой классификации мы постоянно решаем в профессиональной деятельности и в быту, пользуясь субъективными вероятностями событий. Для ответа на этот вопрос проведем на рис. 2 линию уровня 0,8 и точки пересечения этой линии с графиками трех ФП спроектируем на ось ТИ. Решения пропорций дают следующие четыре значения ТИ: 4,18, 4,88, 5,30, 6,00.

Соотнося эти значения с исходной таблицей (процесс перехода из области значений ФП в предметную область называется *дефазификацией*), получаем ответы:

– с уверенностью более 0,8 к классу (х) относятся объекты 1–6;

– с уверенностью более 0,8 к классу (т) относятся объекты 16–18;

– с уверенностью более 0,8 к классу (ВТ) относятся объекты 22–26.

4. Зона значений признака от 5,06 до 5,12 выявила характерную особенность нечетких множеств, а именно сумма значений ФП одного объекта к нескольким НМ может быть больше единицы. Гипотетические объекты со значениями ТИ = 5,06 и ТИ = 5,12 имеют эту особенность: по графикам видно, что сумма значений двух ФП больше 1. В классической теории вероятностей имеет место строгое равенство: $P(A) + P(\neg A) = 1$ (полная группа несовместных событий).

5. В этой же зоне значений признака от 5,06 до 5,12 функция принадлежности объектов к классу «теплолюбивые» двузначная (в общем случае бывает многозначная). В нечетких моделях это типичная ситуация, в ней отражается сложность (противоречивость) реальных явлений. Например, отмеченную двузначность можно трактовать как разные оценки холодовыносливости одних и тех же объектов двумя экспертами. Для принятия решений в подобных ситуациях Л. Заде предложил *принцип обобщения*. В нашем случае этот эвристический принцип можно представить следующим образом: имеется несколько экспертов-НМ, каждому из которых мы одинаково доверяем, но их оценки одного и того же объекта по какому-либо свойству не совпадают. В качестве решения ситуации принимаем мнение того эксперта, у которого оценка максимальна. На рис. 2 принцип обобщения соответствует верхней огибающей двух ФП к классу (т) в зоне значений признака от 5,06 до 5,12.

ВЫВОДЫ

Способ нечеткой классификации малого числа объектов, описанный в данной работе, основан на нескольких эвристических допущениях, которые, разумеется, могут оспариваться. Но эти допущения при помощи теории нечетких множеств позволили выявить дополнительные соотношения между реальными объектами. Результаты анализа соответствуют взглядам специалистов предметной области.

В условиях неприменимости к короткому вариационному ряду статистических способов анализа продемонстрированы возможности теории нечетких множеств для реалистичной оценки принадлежности объектов к классам, для идентификации новых объектов (нечеткого дискриминантного анализа), для формулирования нечетких продукционных правил при построении экспертных систем.

Описанный способ нечеткой классификации малого числа объектов неспецифичен для предметной области (биологии), поэтому может широко

применяться, поскольку шкала отношений – одна из самых распространенных в измерениях физических и технологических параметров.

В заключение автор выражает признательность д-ру биол. наук И. А. Куперману за советы и замечания, высказанные во время работы над статьей.

СПИСОК ЛИТЕРАТУРЫ

1. **Гордеева Н. И.** Опыт ранжирования растений по требовательности к теплу с помощью температурного коэффициента стандартизированного метаболизма: Автореф. дис. ... канд. биол. наук /ЦСБС СО РАН. Новосибирск, 1999.
2. **Загоруйко Н. Г.** Гипотезы компактности и λ-компактности в методах анализа данных // Сибирский журнал индустриальной математики. 1998. **1**, № 1. С. 114.
3. **Заде Л. А.** Понятие лингвистической переменной и его применение к принятию приближенных решений. М.: Мир, 1976.
4. **Дюбуа Д., Прад А.** Теория возможностей. Приложения к представлению знаний в информатике: Пер. с фр. М.: Радио и связь, 1990.
5. **Борисов А. Н., Алексеев А. В., Меркурьева Г. В. и др.** Обработка нечеткой информации в системах принятия решений. М.: Радио и связь, 1989.
6. **Саати Т.** Принятие решений. Метод анализа иерархий: Пер. с англ. М.: Радио и связь, 1993.
7. **Пфанцагль И.** Теория измерений. М.: Мир, 1976.

Центральный сибирский ботанический сад СО РАН,
E-mail: root@botgard.nsk.su

Поступила в редакцию
10 февраля 2000 г.