РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ
АВТОМЕТРИЯ
№ 3 1999

УДК 519.816 : 58.002

В. И. Красинский
(Новосибирск)

ПРИМЕНЕНИЕ ТЕОРИИ ВОЗМОЖНОСТЕЙ ДЛЯ РАНЖИРОВАНИЯ МНОГОЗНАЧНЫХ БОТАНИЧЕСКИХ ОБЪЕКТОВ

Применены теория нечетких множеств и связанная с ней теория возможностей для ранжирования многозначных ботанических объектов (семейств). В живой природе представители большинства таксонов-семейств имеют разные значения признака внутри своего таксона. Такую информацию затруднительно обработать статистическими методами. Вместо распределения вероятностей значений морфологического признака числа лепестков вычисляются распределения мер необходимости и возможности этих значений по двум последовательностям вложенных подмножеств исходного множества – фокальным подмножествам. Фокальные подмножества построены по определению противоположных событий «быть меньше или равно K» (слева) и «быть больше K» (справа), где K – значения признака в интервалах группировки. На этой основе получены диапазоны (вместо конкретных чисел) априорных вероятностей всех значений признака и диапазон среднего значения признака. По мерам необходимости слева и справа вычислены значения функции принадлежности (ФП) объектов к итоговому нечеткому множеству (НМ) как степени многозначности (размытости) объектов. Величина, дополнительная к ФП, названа двусторонним коэффициентом предпочтения, она является гарантированным уровнем доверия к диагнозу объекта по значению признака. Предложен новый эвристический прием нормализации НМ. Вычислен энтропийный коэффициент нечеткости НМ. Предложено использовать максимальные разности значений ФП для прогноза аномальных объектов. Приводятся все исходные данные значений чисел лепестков 102 семейств двудольных растений Сибири и результаты вычислений. Изложенная методика анализа многозначного числового признака может применяться и в других предметных областях, так как показатели нечеткости безразмерны, а операции с нечеткими множествами абстрагированы от физической сущности задачи.

Введение. Математическая проблема классификации ботанических объектов осложняется многозначностью числовых показателей, характеризующих эти объекты, а именно, значение признака может быть разным у растений одного таксона-семейства. В статистике такой случай трактуется как ошибка в исходных данных (нет полной группы несовместных событий). В данной статье анализируется показатель числа лепестков двойного или членов простого околоцветника, кратко он называется числом лепестков. Например, в семействе Saxifragaceae встречаются особи с одним, четырьмя или пятью лепестками. Около 3/4 семейств имеют подобную многознач-

ность, поэтому даже в рассматриваемом одномерном случае объекты невозможно представить как вариационный ряд по значению признака.

Формализацию описания подобных объектов и взаимосвязанные способы классификации можно успешно осуществить при помощи теории нечетких множеств (НМ) и развитой на ее основе теории возможностей, поскольку в них операндами являются значения лингвистических термов типа «больше», «меньше», а также подмножества объектов, для которых установлены отношения порядка.

1. **Постановка задачи и обоснование выбора метода анализа информации.** Одной из задач автоматизации флористических исследований, проводимых в ЦСБС СО РАН под руководством д. б. н. И. М. Красноборова, является создание программы определителя семейств растений Сибири. Информация о 102 семействах задана таблицей признаков: 7 – номинальных, 4 – числовых. Поэтому требуется провести кластерный анализ объектов в многомерном пространстве разнотипных признаков. Практически все признаки многозначные, вследствие этого применение известных программ многомерного статистического анализа СИГАМД, STATISTICA для кластеризации семейств не привело к успеху.

В качестве альтернативного теории НМ способа обработки нечеткой числовой информации можно было бы применить интервальное описание и связанные с ним методы анализа. В этом случае указываются минимальное и максимальное значения признака для каждого объекта (измерения). Внутри интервала принимается равномерное распределение вероятностей значений признака. Однако таблица исходных данных (приложение) по семействам растений имеет более сложную структуру (например, в упомянутом семействе Saxifragaceae возможны значения признака 1, 4, 5, однако значения 2 и 3 невозможны). Подобных «неравномерных» объектов имеется 12 из 102, поэтому если принять для всей задачи интервальное представление признака, то это было бы значительным огрублением исходной информации. Предлагаемый автором способ анализа многозначного признака учитывает исходные данные без всякого упрощения.

Методы интервального анализа развиты пока только для одномерного случая, в них также нет способа представления номинальных признаков. Методика постановки и теория решения более общих так называемых локализационных задач еще только разрабатываются. Опираясь на теорию Демпстера – Шефера, в недавней работе [1] предлагается представлять числовые интервалы как вложенные множества. Таким образом, признается плодотворность этой теории и для интервальных методов анализа, но не представлен пример с результатами ее применения.

Цель настоящей работы – априорное ранжирование объектов по степени нечеткости одного многозначного числового признака. Это способствует уточнению диагноза исследуемого объекта. Предлагаемый для ранжирования способ основан на теории Демпстера – Шефера с уточнениями Дюбуа и Прада и в отличие от интервального метода учитывает все исходные данные без упрощения.

Теория НМ весьма развита в направлении многомерного анализа, дает возможность представить с единых абстрактных позиций разнотипную исходную информацию, ограничения и цели, поэтому полученные в настоящей работе результаты будут применены для решения задачи классификации ботанических объектов по совокупности многозначных разнотипных признаков.

2. **Теоретические предпосылки.** Понятие нечеткого множества и основные операции над ними впервые ввел Л. Заде [2]. Также можно указать, например, на его работы [3–6]. Теория НМ хорошо изложена в [7].

Пусть A и B – нечеткие подмножества универсального множества X, при этом $x \in X$. Это означает, что значения характеристической функции принадлежности (ФП) элемента универсума к нечетким подмножествам A, B не равны только числам 0 и 1, а могут быть любыми действительными числами из этого диапазона. Пусть также рассматривается событие $(x \in A \cap x \in B)$. Для измерения степени этой совместной принадлежности А. Демпстер [8] ввел понятия *неаддитивных нижних и верхних вероятностей* P_1, P_2. Значение вероятности события $(x \in A \cap x \in B)$ находится между ними. Таким образом, вероятность многозначного события заменяется некоторым ее диапазоном.

В дальнейшем изложении имеется в виду, что формулировка некоторого события по отношению к множеству отделяет соответствующее подмножество его элементов, поэтому не различаются строго термины «подмножество» и «событие».

Теорию возможностей также сформулировал Л. Заде [3, 6]. В этой теории неопределенность некоторого события описывается степенями одновременной возможности события и возможности противоположного события.

Мерами возможности по Заде [6] называются функции П (иногда обозначают Pos от "possibility") множеств такие, что

$$\forall A, B \subseteq X, \qquad \Pi(A \cup B) = \max(\Pi(A), \Pi(B)), \tag{2.1}$$

при этом нет предположения о том, что A и B – непересекающиеся множества (несовместные события).

Обозначим посредством $\neg A$ событие, противоположное событию A.

Другой класс функций множеств, называемых *мерами необходимости* и обозначаемых N ("necessity"), удовлетворяет аксиоме

$$\forall A, B \subseteq X, \qquad N(A \cap B) = \min(N(A), N(B)). \tag{2.2}$$

Постулируется, что значения $\Pi(A)$ и $N(A)$ лежат в диапазоне $0 \div 1$. Аксиомы (2.1) и (2.2) непротиворечивы только тогда, когда имеют место соотношения:

$$\forall A, \ \Pi(A) = 1 - N(\neg A), \tag{2.3}$$

$$\forall A, \ N(A) = 1 - \Pi(\neg A). \tag{2.4}$$

Положив в (2.4) $\Pi(\neg A) = 0$, получим постулат модальной логики: «событие необходимо, когда противоположное событие невозможно».

Имеют место соотношения:

$\max(\Pi(A), \Pi(\neg A)) = 1$ (из двух противоположных событий одно, безусловно, возможно);

$\min(N(A), N(\neg A)) = 0$ (исключена одновременная необходимость двух противоположных событий);

$$\Pi(A) \geq N(A); \qquad N(A) > 0 \Rightarrow \Pi(A) = 1; \qquad \Pi(A) < 1 \Rightarrow N(A) = 0;$$

$$\Pi(A) + \Pi(\neg A) \geq 1; \qquad N(A) + N(\neg A) \leq 1.$$

В классической теории вероятностей вместо двух последних соотношений имеет место $P(A)+P(\neg A)=1$. Степень необходимости отражает свидетельства в пользу события, а степень возможности – свидетельства против этого события.

Если множество X конечно, то меру возможности события A можно определить по ее значениям на одноточечных подмножествах X:

$$\Pi(A)=\sup\{\pi(x)\,|\,x\in A\},$$

где $\pi(x)=\Pi(\{x\})$; π есть отображение из X в $[0, 1]$, называемое функцией распределения возможностей. Оно является нормальным в смысле $\exists x,\ \pi(x)=1$, поскольку $\Pi(X)=1$.

Функцию распределения необходимости также можно построить по функции распределения возможности:

$$N(A)=\inf\{1-\pi(x)\,|\,x\notin A\}.$$

Связь между мерами необходимости и возможности, с одной стороны, и неаддитивными вероятностями Демпстера – с другой, показал Г. Шефер [9]: «нижняя вероятность P_1» – это мера необходимости $N(A)$, а «верхняя вероятность P_2» – это мера возможности $\Pi(A)$.

Распределением уверенности (необходимости) по Шеферу называется функция ("belief function") Bel: $2^X \to [0,1]$, обладающая следующими свойствами на конечном множестве событий A_i:

$$\mathrm{Bel}(\varnothing)=0,\quad \mathrm{Bel}(A_i)\le 1,\quad \sum \mathrm{Bel}(A_i)=1 \text{ по всем } A_i\subseteq X,$$

т. е. требование полноты группы несовместных событий заменяется распределением единичной «массы уверенности» на все возможные события. Величина $\mathrm{Bel}(A_i)$ выражает степень необходимости события A_i (четкого подмножества). Если $\mathrm{Bel}(A_i)>0$, то подмножество A_i называется *фокальным элементом* (ФЭ) распределения уверенности на множестве X.

Переход от мер уверенности (необходимости) к мерам вероятности составляет существо теории Демпстера – Шефера. Г. Шефером доказано [9], что меры P_1 и P_2 являются соответственно мерами необходимости и возможности тогда и только тогда, когда ФЭ образуют *последовательность вложенных подмножеств*. Как следствие этого доказательства величина $\mathrm{Bel}(A_i)$ есть *вероятность совокупности элементарных событий* (объектов), составляющих ФЭ, причем не учитывается распределение величины $\mathrm{Bel}(A_i)$ по этим событиям. Иначе говоря, ФЭ – *это класс элементарных событий, принадлежащих универсуму X и объединенных одинаковой мерой необходимости.*

Понятие ФЭ и доказательство Шефера позволяют формализовать многозначность (субъективизм) мнений экспертов или многозначность свойств объектов, возникающую вследствие сложности моделируемых явлений.

Переобозначим для краткости $m(A_i)=\mathrm{Bel}(A_i)$. Итак, исходное множество разбивается на непустые, попарно различные четкие подмножества $A_1, A_2, \ldots, A_r$. Если при помощи формулировок r событий построить эти подмножества как *согласованные*, т. е. обеспечить вложенность

$$A_1\subseteq A_2\subseteq\ldots\subseteq A_r, \tag{2.5}$$

то количества отнесенных к ним элементарных объектов (мощности этих подмножеств) определят частотные вероятности приращений достоверности универсума по вложенным событиям:

$$m(A_i) = (\mathrm{Card}A_i - \mathrm{Card}A_{i-1})/\mathrm{Card}X, \qquad i = 1,\ldots,r. \tag{2.6}$$

Имеется в виду, что мощность пустого множества $\mathrm{Card}\varnothing = 0$.

Подмножества $A_1,\ldots,A_r$ называются *фокальными подмножествами* (*списками*). Условие $\sum m(A_i) = 1$ будет выполнено, так как рассматривается только счетное универсальное множество X, и каждый элементарный объект $x \in X$ окажется отнесенным только к минимальному по вложению ФЭ.

В совокупности выражений (2.5) и (2.6) состоит алгоритм получения нижней оценки приращений необходимости события A_i (аналогично вычислению эмпирической функции распределения вероятностей).

В согласованности ФЭ проявляется необходимость (мера P_1) события A_{i+1} по отношению к событию A_i, а именно, из A_i следует A_{i+1}.

Важное положение доказано П. Дюбуа и Г. Прадом в [10]: если имеет место согласованность ФЭ, то функция распределения возможностей π, связанная с P_1 и P_2, определяется следующим образом:

$$\forall x,\ \pi(x) = P_2(\{x\}) = \sum_{j=i} m(A_j), \text{ если } x \in A_i \ \text{ и } \ x \notin A_{i-1}.$$

Итак, нижнюю $P_1(A)$ и верхнюю $P_2(A)$ оценки вероятности события A можно получить из эмпирического распределения ФЭ:

$$P_1(A) = \sum m(A_i) N_{A_i}(A), \qquad P_2(A) = \sum m(A_i) \Pi_{A_i}(A). \tag{2.7}$$

Выражения (2.7) дают алгоритм вычисления неаддитивных вероятностей Демпстера неопределенного события: значение $P_1(A)$ вычисляется по всем ФЭ, которые делают необходимым появление события A (или влекут за собой событие A); значение $P_2(A)$ вычисляется по тем ФЭ, которые оставляют возможным появление события A. Если ФЭ будут состоять только из элементарных (а значит, и несовместных) событий, то получится стандартная гистограмма и возврат к вероятностной мере уже как к частному случаю: $\forall A,\ P_1(A) = P_2(A) = P(A)$.

А. Демпстер в [8] строго определил понятия нижних и верхних математических ожиданий для конечного множества с помощью интеграла Лебега – Стилтьеса в рамках мер необходимости и возможности:

$$E_1(f) = \int r dN(\{x \mid f(x) \le r\}), \qquad E_2(f) = \int r d\Pi(\{x \mid f(x) \le r\}),$$

здесь $f(x)$ – функция, определенная на X и принимающая значения из множества действительных чисел. Таким образом вычисляется диапазон значений математического ожидания функции от нечеткого аргумента, вместо одного значения для классического случая.

В [11] отмечается, что вероятностные меры отражают точные и дифференцированные знания, а меры возможности и необходимости – неточные, но связанные каким-либо отношением знания.

Нечеткое множество можно определить совокупностью обычных множеств уровня α (α-срезами F_α) следующим образом:

$$\forall x,\ \mu_F(x) = \sup\{\alpha \mid x \in F_\alpha\}.$$

Здесь $\mu_F(x)$ – функция принадлежности элементов (объектов) к нечеткому множеству [7].

В работах [12, 13] показано, что НМ можно рассматривать как «след» меры возможности на одноточечных множествах в X:

$$\forall \Pi,\ \exists F \in [0,1]^X, \quad \forall x \in X, \quad \Pi(\{x\}) = \pi(x) = \mu_F(x).$$

С другой стороны, задание НМ достаточно для описания функции распределения возможностей при условии, что это НМ нормально, т. е. $\exists x,\ \mu_F(x) = 1$.

В [13] показано, что если функция распределения возможностей определяется по относительным частотам согласованных ФЭ, то эти ФЭ образуют семейство четких подмножеств уровня (α-срезов) некоторого нечеткого множества F: $A_i = F_{\alpha i}$, где $\alpha_j = \sum m(A_j)$, $j = 1, \ldots, i$ (j – индекс ФЭ).

В работе [12] обосновывается уточненное по сравнению с $\pi(x)$ вычисление ФП элементов универсального множества X к нечеткому множеству:

$$\forall x,\ \mu_F(x) = \sum m(A_j) T_j(x), \tag{2.8}$$

где $T_j(x)$ – характеристическая функция принадлежности элемента x к четкому подмножеству A_j. Уточнение (2.8) очень важно, оно будет использоваться в последующих вычислениях, а именно, $\mu_F(x) \le \pi(x)$, и поэтому значения ФП, вычисленные по (2.8), представляют собой нижние оценки вероятности (необходимости) принадлежности элементарных объектов x к НМ F.

Задавая разные уровни α-срезов нечеткого множества, можно получать четкие подмножества (списки) элементарных объектов с *гарантированной вероятностью* $1 - \alpha$, интерпретируемые как «доверительные множества» подобно доверительным интервалам в статистике, поскольку функция принадлежности к НМ вычисляется по мерам необходимости ФЭ $m(A_j)$.

Именно в этом смысле уровня доверия в разд. 4 получен итоговый коэффициент $K2$ предпочтения диагноза элементарного объекта по измеренному в эксперименте значению признака.

3. **Формулировка задачи в предметной области.** Вышеприведенные теоретические положения позволяют перейти к анализу неоднозначных экспериментальных данных. Для этого необходимо определить множественные отношения порядка для событий (вложенных четких подмножеств) с целью построения ФЭ и вычисления нижней и верхней оценок вероятностей значений признака и построения ФП объектов к НМ.

Таблица исходных данных (приложение) является двоичной матрицей инцидентности 102 строк объектов-семейств и 7 колонок квантованных значений признака – числа лепестков. Колонки соответствуют значениям 0, 1, 2, 3, 4, 5, >5. Примерно в 3/4 строк таблицы единица установлена более чем в одной колонке. Наиболее неоднозначными являются семейства Betulaceae – m., Chenopodiaceae (All), Ulmaceae: у них имеется по пять единиц в строке (из семи возможных).

При подобных исходных данных отношение порядка для построения вложенных подмножеств устанавливается определением события «быть

меньше K или равно K», где K – целые числа $\{0, 1, 2, 3, 4, 5, 99\}$. Число 99 достаточно для данной задачи как максимальное значение признака числа лепестков. Множественное отношение вложенности при этом обеспечено: если объект имест, скажем, 2 или меньше лепестков, то из этого следует «3 или меньше», «4 или меньше» и т. д. Таким образом формируем семь четких ФЭ $A_1 - A_7$. Относительную разность мощностей соседних ФЭ в соответствии с (2.6) обозначим $ml(K)$ – разность частот слева. Интересующими нас событиями, характеризующимися необходимостью, или нижней оценкой вероятности P_1, является «равно K или меньше K», а также «равно K». Тогда нижняя оценка P_1 для события, например, «3 или меньше», соответствующего ФЭ A_4, определится относительной суммой $ml(1) + ml(2) + ml(3)$, т. е. кумулянтой тех событий, которые делают необходимым событие A_4.

Постулатом противоположного по отношению к событию «быть меньше K или равно K» является событие «быть больше K». По этому событию построим свою фокальную последовательность – четкие подмножества исходного множества $B_1 - B_7$. Эти подмножества требуются для вычисления верхних вероятностей P_2. В подмножествах $B_1 - B_7$ также соблюдается строгая вложенность, поскольку очевидно, что из условия «больше 5» следует «больше 4», «больше 3» и т. д. Аналогично и по остальным числовым сравнениям ФЭ B_k. Их относительные разности мощности обозначим $mr(K)$ – разность частот справа – также в соответствии с (2.6).

Нижняя (гарантированная) оценка вероятностей событий «равно K» будет $P_1(K) = ml(K)$, поскольку ФЭ были построены по принципу вложенности. Аналогично частоты $mr(K)$ дают необходимость событий «равно K», но полученную по вложенным ФЭ справа, т. е. как событий, противоположных событиям «меньше K или равно K». Поэтому в соответствии с выражением (2.3) можно вычислить верхние вероятности (возможности) событий «равно K»: $P_2(K) = 1 - mr(K)$.

Теперь по мерам необходимости $ml(K)$ и $mr(K)$ в соответствии с (2.8) можно построить функции принадлежности объектов к двум нечетким множествам $\mu_L(x)$ и $\mu_R(x)$, соответственно назовем их «нечеткость слева» и «нечеткость справа». Эти два НМ позволяют наилучшим образом в смысле максимина учесть многозначность объектов.

Приведенные в разд. 2 соотношения с ФП к нечеткому множеству предполагают нормальность этого множества, т. е. $\exists x$, $\mu(x) = 1$, иначе, среди исходных объектов должен быть хотя бы один такой, который имеет все значения признака (максимально нечеткий объект). В таблице исходных данных таких объектов нет, поэтому требуется способ нормализации функций $\mu_L(x)$ и $\mu_R(x)$. Для этой цели обычно прибегают к масштабированию максимальным значением.

В настоящей работе применяется другой прием: к списку объектов добавлен один искусственный, имеющий инцидентность со всеми значениями признака. Этим самым нечеткое множество нормализуется. Один объект в дополнение к 102 незначительно изменяет частоты ФЭ. При этом сохранение практически исходных значений ФП $\mu_L(x)$ и $\mu_R(x)$ важно для последующих операций совместного анализа нескольких нечетких числовых и нечисловых показателей, разные же масштабы функций принадлежности затруднят многомерный анализ.

4. **Числовые результаты.** 4.1. *Диапазоны вероятностей сравнительных значений признака*. Обозначим имена ФЭ A_k как LE K ("less or equivalent then

Таблица 1

Фокальные элементы слева

Фокальные элементы	LE 0	LE 1	LE 2	LE 3	LE 4	LE 5	LE 99	Сумма
Количество	13	2	3	12	32	37	4	103
$ml(K)$	0,13	0,02	0,03	0,12	0,31	0,35	0,04	1,0
Кумулянта $ml(K)$	0,13	0,15	0,18	0,30	0,61	0,96	1,0	

Таблица 2

Фокальные элементы справа

Фокальные элементы	G 0	G 1	G 2	G 3	G 4	G 5	G 99	Сумма
Количество	4	1	0	3	16	60	19	103
$mr(K)$	0,039	0,01	0,0	0,029	0,155	0,583	0,184	1,0
Кумулянта $mr(K)$	1,0	0,961	0,951	0,951	0,922	0,767	0,184	

K"), а имена ФЭ B_k как G K ("greather then K") в соответствии с их определением в разд. 3.

Компьютерное преобразование исходной информации о 102 + 1 объектах (из приложения) в ФЭ по алгоритму, изложенному в разд. 2, привело к следующим результатам, представленным в табл. 1, 2.

Используя эти таблицы и соотношения, представленные в разд. 2, вычислим нижние и верхние априорные вероятности значений чисел лепестков. А именно, поскольку значения $ml(K)$ из табл. 1 учитывают необходимость равенства числу K, принимаем эти значения как нижние (гарантированные) вероятности $\underline{P}_1$ соответствующего числа лепестков. Верхние вероятности значений числа лепестков получаются из табл. 2 вычислением $P_2(K) = 1 - mr(K)$, что было показано в разд. 3. Сводим результаты в табл. 3.

Например, априорная вероятность события «объект имеет число лепестков 3» находится в диапазоне от 0,12 до 0,971. Из сопоставления табл. 1–3 видно, что два наиболее многочисленных ФЭ как слева, так и справа образуют события, соответствующие числам лепестков 4 и 5. Поэтому можно предсказать, что гарантированное нечеткое среднее значение числа лепестков находится в интервале от 4 до 5 (в дальнейшем это подтвердится расчетом). Наиболее узкий диапазон нижней и верхней вероятностей соответствует этому же интервалу 4 и 5 лепестков, что также подтверждает наличие моды в этих интервалах группирования.

Таблица 3

Диапазон вероятностей значений признака

Значение признака	= 0	= 1	= 2	= 3	= 4	= 5	> 5
Верхняя P_2	0,961	0,99	1,0	0,971	0,845	0,417	0,816
Нижняя P_1	0,13	0,02	0,03	0,12	0,31	0,35	0,04

На основании качественного анализа табл. 1–3 можно предложить разбиение исходного множества семейств на три класса с алгоритмом группирования по числу лепестков: «равно или меньше 3», «равно 4 или 5», «больше 5». Эти три класса образуют другое НМ, с меньшей нечеткостью, чем исходное, но в настоящей работе оно не анализируется.

Для технических и технологических задач по таблицам, составленным аналогично табл. 1–3, группируя интервалы (суммируя $ml(K)$, $mr(K)$), можно получать меры необходимости и возможности этих расширенных диапазонов, переходя к задачам *нечеткого допускового контроля в технологиях*.

4.2. *Коэффициент предпочтения для определения семейств*. Вычислим значения функций принадлежности объектов к двум НМ:

$\mu_L(x)$ – к нечеткому множеству L, т. е. определим «степень нечеткости» слева значения признака числа лепестков для каждого объекта-семейства по указанному в разд. 2 алгоритму (выражение (2.8)):

$$\forall x,\ \mu_L(x) = \sum ml(K_j)T_j(x);$$

$\mu_R(x)$ – к нечеткому множеству R, т. е. аналогично определим «степень нечеткости» этого признака справа для каждого объекта:

$$\forall x,\ \mu_R(x) = \sum mr(K_j)T_j(x).$$

Вычисленные значения ФП $\mu_L(x)$ и $\mu_R(x)$ находятся в пределах от 0 до 1 и удовлетворяют свойствам так называемого изотонического показателя [14], или «потенциала», «ранга» данного объекта в рассматриваемой совокупности объектов. Этот показатель не сохраняет структуру свойств объектов (безразлично, из принадлежности к каким ФЭ набран «общий вес» каждого объекта) и учитывает только уровень значений свойств $ml(K_j)$ или $mr(K_j)$. Равные значения изотонического показателя образуют изокванты Хельвига, которые объединяют объекты, равноудаленные от начала координат в смысле метрики расстояния "city-block" («манхэттенское»). Это частный случай метрики Минковского, оперирующей функцией от модуля разности многомерных расстояний.

Таким образом, α-срезы НМ, образованного по ФЭ, являются изоквантами Хельвига, что дает теоретическое обоснование для классификации объектов, представленных матрицей расстояния $d_{ij} = |\mu(x_i) - \mu(x_j)|$. Известные программные пакеты СИГАМД, STATISTICA имеют соответствующие режимы.

Итак, имеются два нечетких множества, в равной степени правомерные представлять исходные многозначные объекты. Двойственность появилась из-за того, что признак числовой, поэтому было установлено два алгоритма получения вложенных подмножеств – «быть меньше K или равным K» и «быть больше K». Суперпозицию двух НМ $\mu_L(x)$ и $\mu_R(x)$ в итоговое НМ $\mu_V(x)$ следует произвести так, чтобы максимально учесть исходную информацию. Множества L и R были построены по необходимым (гарантированным) априорным вероятностям значений признака, поэтому есть основание применить к ним операцию объединения, чтобы не уменьшать эти гарантированные вероятности. В теории НМ этот эвристический прием соответствует принципу обобщения Заде:

$$\mu_V(x) = \mu_{L \cup R}(x) = \max\{\mu_L(x),\ \mu_R(x)\},\quad x \in X.$$

Полученное для каждого объекта значение ФП $\mu_V(x)$ *отражает гарантированную степень его принадлежности к НМ, т. е. его нечеткость, с учетом всей таблицы исходных данных. Поэтому чем больше значение* $\mu_V(x)$, *тем хуже распознается объект по признаку числа лепестков.*

Дополнение к нечеткому множеству V (множество $\neg V$) показывает степень различимости (специфичности) объектов по анализируемому показателю, поэтому назовем его функцию принадлежности двусторонним коэффициентом предпочтения диагноза $K2(x) = 1 - \mu_V(x)$. Его теоретический смысл как уровня $1 - \alpha$ доверительных множеств показан в разд. 2.

Для практического применения выдан список всех 102 семейств, упорядоченный по убыванию коэффициента $K2$. Этот список, совместно с исходными данными, предоставляет ключ для отделения подмножества семейств, имеющих значение $K2$ не меньше заданного, т. е. априорно определены объекты, имеющие изокванту $1 - \alpha$ не менее заданной. Для примера приводим подсписок семейств, имеющих 1 лепесток (табл. 4). Таких семейств оказалось шесть. Поэтому, если учитывать только таблицу исходных данных, то исследуемое растение с одним лепестком следует равновероятно отнести к любому из шести семейств – кандидатов на диагноз. Если же использовать полученный коэффициент $K2$, то шесть кандидатов априорно ранжированы, притом весьма сильно.

Качественный вывод совпадает с рангом $K2$: на последнем месте находится самый «размытый» (сложный) объект – возможны 5 из 7 значений признака, на первом месте – самый «определенный» объект. Как вычислено ниже в разд. 4.3, среднее значение нечеткого признака равно «от 4 до 5», поэтому четыре объекта с этими возможными значениями находятся в списке ниже, чем первые два, более специфичные по этому показателю.

Итак, если учитывать только один признак числа лепестков и у предъявленного объекта это число равно 1, то диагнозом семейства является "Cannabaceae – f." с уровнем доверия 0,855. Вторым кандидатом является "Betulaceae – f." с уровнем доверия 0,709 и т. д.

Вычисленная ФП $\mu_{1\cdot}(x)$ по числу лепестков будет нужна в многомерном анализе на основе математического аппарата операций с НМ.

Другой практический результат использования ранжированного по значению коэффициента $K2$ списка состоит в прогнозе объектов с наиболее

Таблица 4

Семейства	$K2$	= 0	= 1	= 2	= 3	= 4	= 5	> 5
Cannabaceae – f.	0,855	1	1	0	0	0	0	0
Betulaceae – f.	0,709	1	1	1	1	0	0	0
Betulaceae – m.	0,398	1	1	1	1	1	0	0
Cannabaceae – m.	0,368	1	1	0	0	0	1	0
Saxifragaceae	0,252	0	1	0	0	1	1	0
Chenopodiaceae (All)	0,165	0	1	1	1	1	1	0

вероятными ошибками в исходных данных. Это те семейства, между которыми наблюдаются наибольшие разности значений $K2$. Ошибки предполагаются общие – как технических работников, так и специалистов. Если для примера рассмотреть только табл. 4, то в ней кандидатами на проверку исходных данных являются второй и третий объекты – между ними наибольшее изменение $K2$. Список таких «подозрительных» семейств предоставлен пользователям-ботаникам.

Аналогично максимальное *изменение* значения коэффициента предпочтения в ранжированном списке объектов может быть использовано для выявления аномальных объектов и определения границ классов в кластерном анализе, поскольку на этот абстрактный показатель для каждого объекта влияет вся таблица исходных данных. Классы здесь следует понимать как классы нечеткости, т. е. объединяются объекты с гарантированным значением $(1-\alpha)$ надежности диагноза объектов.

4.3. *Обобщенные характеристики многозначного признака.* 4.3.1. Среднее значение. Применим к ФЭ обычную в теории вероятностей операцию вычисления математического ожидания дискретной случайной величины: $M[X]=\sum_i x_i p_i$ (x_i – значения интервалов гистограммы).

Применяя упомянутые в разд. 2 обоснования вычисления интегралов по мерам необходимости, найдем нижнюю границу слева среднего значения числа лепестков по вероятностям $ml(K)$ из табл. 1, осуществив только соответствующее предметной области присвоение значения 7 самому правому ФЭ, после чего получим $M[\mathrm{LE}]=3{,}71$, округленно $M_{\mathrm{LE}}=4$. Аналогично по вероятностям $mr(K)$ из табл. 2 получим нижнюю границу справа среднего значения: $M[\mathrm{G}]=4{,}92$, округленно $M_{\mathrm{G}}=5$.

Итак, получено нечеткое среднее значение показателя по исходной информации: «семейства имеют в среднем от 4 до 5 лепестков». Такая интервальная оценка, вместо одного числа, является следствием многозначности исходных данных. Это правило может применяться для разделения семейств на три класса, о чем уже сообщалось выше.

4.3.2. Коэффициент нечеткости многозначного признака. Для оценки степени нечеткости, или «размытости» исходных данных (нечеткого множества F), применяются разные коэффициенты. Их аксиоматика изложена в [7]. Одна из этих аксиом очень характерна: коэффициент нечеткости $f(F)$ максимален тогда и только тогда, когда $\mu_{F'}(x)=0{,}5$. Это соответствует наиболее неопределенному значению принадлежности (ситуация «не знаю»). Именно это значение характеристической ФП используется в трехзначной логике Лукашевича. Другая аксиома требует, чтобы было $f(F)=0$ тогда, когда F – обычное четкое подмножество множества X.

Одной из функций, удовлетворяющей аксиоматике для $f(F)$, является известная в теории информации функция Шеннона:

$$S(p)=-p\ln(p)-(1-p)\ln(1-p),$$

где p – вероятность события.

Максимальное значение $S_{\max}(p)=0{,}6931$ функция Шеннона принимает при аргументе $p=0{,}5$. Это соответствует наибольшей неопределенности события и называется «1 нат информации». Если в функции Шеннона применить двоичные логарифмы, то ее максимальное значение равно 1 и называется «1 бит информации». Таким образом, 1 бит = 0,6931 нат.

На основе этой функции в работе [15] обосновывается понятие логарифмической энтропии нечеткого множества:

$$Z(F) = k\sum S(\mu_F(x_i)),$$

где сумма вычисляется по всем элементам нечеткого множества, а k – положительная константа. Термин «энтропия» удачно подходит для характеристики степени неопределенности НМ. Обратим внимание, что для обычного (четкого) множества энтропия $Z(F) = 0$, поскольку значениями аргумента функции $S(\mu_F(x_i))$ могут быть только 1 или 0.

Положим масштабирующий множитель $k = 1$ и вычислим логарифмическую энтропию по всем 103 элементам для нечетких множеств L, R, V: $Z(L) = 60{,}9133$ нат, $Z(R) = 54{,}226$ нат, $Z(V) = 58{,}9045$ нат. Эти величины малоинформативны, но оценим их относительные значения. Умножив число объектов на величину $S_{max}(p)$, получим «потенциальный максимум» энтропийности (размытости) НМ, состоящего из 103 объектов:

$$Z_{max}(F) = 103 \cdot 0{,}6931 \text{ нат} = 71{,}3893 \text{ нат}.$$

Введем относительный энтропийный коэффициент размытости нечеткого множества $f(F) = (kZ(F))/(kZ_{max}(F))$, который не зависит от основания логарифмов, масштаба k и безразмерен. Вычислим для нечетких множеств L, R, V этот коэффициент:

$$f(L) = 60{,}9133/71{,}3893 = 0{,}8533, \qquad f(R) = 54{,}226/71{,}3893 = 0{,}7596,$$

$$f(V) = 58{,}9045/71{,}3893 = 0{,}8251.$$

Поскольку полученные величины коэффициента размытости близки к 1, можно сделать вывод, что исходные данные «сильно неопределенны», хотя в дальнейшем необходимо сравнивать значения этого коэффициента для разных признаков, а именно, чем больше значение $f(F)$, тем хуже применимы стандартные процедуры статистической обработки и хуже качество диагностики любыми методами. Этот коэффициент, наряду с различными коэффициентами информативности, следует учитывать при выборе признаков для классификации объектов.

ВЫВОДЫ

Показано применение теории возможностей для ранжирования многозначных ботанических объектов. Получены диапазоны априорных вероятностей значений количественного многозначного признака (числа лепестков). Получен интервал среднего значения этого признака.

Для практического использования в определителях семейств пользователям-ботаникам предложен коэффициент предпочтения, представляющий собой гарантированный уровень доверия к диагнозу семейства по числу лепестков предъявленного образца.

Предложено использовать максимальные изменения функции принадлежности к нечеткому множеству ранжированного по значениям этой функции списка объектов как границы классов качества диагноза объектов.

Вычисленные характеристики нечеткого множества инвариантны к физической сущности объектов и их свойств, поэтому операции с этими характеристиками удобны в многомерном анализе, в том числе с применением известных программных комплексов.

Изложенная методика обработки многозначных числовых признаков может применяться при решении задач ранжирования (классификации) объектов в различных предметных областях.

В заключение автор считает своим долгом выразить признательность профессорам д. т. н. В. В. Губареву, д. т. н. В. З. Манусову, д. б. н. И. А. Куперману, д. б. н. Л. И. Малышеву; к. т. н. А. Л. Осипову, принявшим заинтересованное участие в обсуждении статьи и сделавшим ценные замечания.

ПРИЛОЖЕНИЕ

Исходные данные по числу лепестков и коэффициент *K*2 надежности диагноза семейств двудольных растений Сибири

Семейство	*K*2	= 0	= 1	= 2	= 3	= 4	= 5	> 5
Aristolochiaceae	0,883	0	0	0	1	0	0	0
Empetraceae	0.883	0	0	0	1	0	0	0
Callitrichaceae	0,874	1	0	0	0	0	0	0
Euphorbiaceae	0.874	1	0	0	0	0	0	0
Hippuridaceae	0,874	1	0	0	0	0	0	0
Salicaceae	0.874	1	0	0	0	0	0	0
Cannabaceae – f.	0.855	1	1	0	0	0	0	0
Berberidaceae	0.816	0	0	0	0	0	0	1
Ceratophyllaceae	0.816	0	0	0	0	0	0	1
Menispermaceae	0.816	0	0	0	0	0	0	1
Resedaceae	0.816	0	0	0	0	0	0	1
Corylaceae – f.	0.777	1	0	0	0	0	0	1
Corylaceae – m.	0.777	1	0	0	0	0	0	1
Betulaceae – f.	0.709	1	1	1	1	0	0	0
Brassicaceae	0,689	0	0	0	0	1	0	0
Celastraceae	0,689	0	0	0	0	1	0	0
Cornaceae	0,689	0	0	0	0	1	0	0
Fumariaceae	0,689	0	0	0	0	1	0	0
Halorrhagidaceae	0.689	0	0	0	0	1	0	0
Hypecoaceae	0,689	0	0	0	0	1	0	0

Семейство	*K*2	= 0	= 1	= 2	= 3	= 4	= 5	> 5
Orobanchaceae	0,689	0	0	0	0	1	0	0
Papaveraceae	0,689	0	0	0	0	1	0	0
Plantaginaceae	0,689	0	0	0	0	1	0	0
Rhamnaceae (Rhamnus)	0,689	0	0	0	0	1	0	0
Trapaceae	0,689	0	0	0	0	1	0	0
Elaegnaceae	0,660	0	0	1	0	1	0	0
Onagraceae	0,660	0	0	1	0	1	0	0
Fagaceae	0,650	0	0	0	0	1	0	1
Lythraceae	0,650	0	0	0	0	1	0	1
Elatinaceae	0,572	0	0	0	1	1	0	0
Urticaceae	0,572	0	0	0	1	1	0	0
Aizoaceae	0,417	0	0	0	0	0	1	0
Apiaceae (Umbelliferae)	0,417	0	0	0	0	0	1	0
Apocinaceae	0,417	0	0	0	0	0	1	0
Asclepiadaceae	0,417	0	0	0	0	0	1	0
Balsaminaceae	0,417	0	0	0	0	0	1	0
Bieberschteiniaceae	0,417	0	0	0	0	0	1	0
Boraginaceae	0,417	0	0	0	0	0	1	0
Campanulaceae	0,417	0	0	0	0	0	1	0
Cistaceae	0,417	0	0	0	0	0	1	0
Convolvulaceae	0,417	0	0	0	0	0	1	0
Diapensiaceae	0,417	0	0	0	0	0	1	0
Droseraceae	0,417	0	0	0	0	0	1	0
Fabaceae	0,417	0	0	0	0	0	1	0
Geraniaceae	0,417	0	0	0	0	0	1	0
Grossulariaceae	0,417	0	0	0	0	0	1	0
Hydrophyllaceae	0,417	0	0	0	0	0	1	0
Hypericaceae	0,417	0	0	0	0	0	1	0
Lentibulariaceae	0,417	0	0	0	0	0	1	0
Limoniaceae	0,417	0	0	0	0	0	1	0

Семейство	*K*2	= 0	= 1	= 2	= 3	= 4	= 5	> 5
Lobeliaceae	0,417	0	0	0	0	0	1	0
Menyanthaceae	0,417	0	0	0	0	0	1	0
Nitrariaceae	0,417	0	0	0	0	0	1	0
Oxalidaceae	0.417	0	0	0	0	0	1	0
Parnassiaceae	0,417	0	0	0	0	0	1	0
Plumbaginaceae	0,417	0	0	0	0	0	1	0
Polemoniaceae	0,417	0	0	0	0	0	1	0
Portulacaceae	0,417	0	0	0	0	0	1	0
Pyrolaceae	0,417	0	0	0	0	0	1	0
Santalaceae	0,417	0	0	0	0	0	1	0
Solanaceae	0,417	0	0	0	0	0	1	0
Tiliaceae	0,417	0	0	0	0	0	1	0
Violaceae	0,417	0	0	0	0	0	1	0
Aceraceae	0,417	0	0	0	0	0	1	0
Betulaceae – m.	0,398	1	1	1	1	1	0	0
Polygalaceae	0,388	0	0	0	1	0	1	0
Malvaceae	0,378	1	0	0	0	0	1	0
Cannabaceae – m.	0,368	1	1	0	0	0	1	0
Adoxaceae	0,262	0	0	0	0	1	1	0
Caprifoliaceae	0,262	0	0	0	0	1	1	0
Caryophyllaceae (All)	0,262	0	0	0	0	1	1	0
Cuscutaceae	0,262	0	0	0	0	1	1	0
Dipsacaceae	0,262	0	0	0	0	1	1	0
Ericaceae	0,262	0	0	0	0	1	1	0
Frankeniaceae	0,262	0	0	0	0	1	1	0
Lamiaceae	0,262	0	0	0	0	1	1	0
Linaceae	0,262	0	0	0	0	1	1	0
Monotropaceae	0,262	0	0	0	0	1	1	0
Rhamnaceae (Frangula)	0,262	0	0	0	0	1	1	0
Rosaceae (All)	0,262	0	0	0	0	1	1	0

Семейство	*K*2	= 0	= 1	= 2	= 3	= 4	= 5	> 5
Scrophulariaceae	0,262	0	0	0	0	1	1	0
Tamaricaceae (Myricaria)	0,262	0	0	0	0	1	1	0
Vacciniaceae	0,262	0	0	0	0	1	1	0
Verbenaceae	0,262	0	0	0	0	1	1	0
Saxifragaceae	0,252	0	1	0	0	1	1	0
Cucurbitaceae	0,233	0	0	0	0	0	1	1
Nymphaeaceae	0,233	0	0	0	0	0	1	1
Paeoniaceae	0,233	0	0	0	0	0	1	1
Primulaceae	0,233	0	0	0	0	0	1	1
Amaranthaceae	0,213	0	0	0	1	1	1	0
Asteraceae	0,213	0	0	0	1	1	1	0
Rubiaceae	0,213	0	0	0	1	1	1	0
Rutaceae	0,213	0	0	0	1	1	1	0
Valerianaceae	0,213	0	0	0	1	1	1	0
Zygophyllaceae	0,213	0	0	0	1	1	1	0
Chenopodiaceae (All)	0,165	0	1	1	1	1	1	0
Crassulaceae (All)	0,078	0	0	0	0	1	1	1
Gentianaceae	0,078	0	0	0	0	1	1	1
Ranunculacea (All)	0,078	0	0	0	0	1	1	1
Thymelaeaceae	0,078	0	0	0	0	1	1	1
Polygonaceae (All)	0,049	0	0	0	1	1	1	1
Ulmaceae	0,010	1	0	0	1	1	1	1

Примечание. После ответа на вопрос относительно предъявленного растения, сколько имеется лепестков, необходимо проследить по соответствующей колонке сверху вниз и в качестве наиболее вероятного кандидата принять первое семейство, у которого встретится единица. Если у нижеследующих семейств с таким же числом лепестков окажется одинаковое с первым значение *K*2, то предпочтения нет, т. е. имеются несколько кандидатов с одинаковой надежностью диагноза.

СПИСОК ЛИТЕРАТУРЫ

1. **Rocha L. R., Kreinovich V., Kearfott R. B.** Computing uncertainty in interval based sets // Applications of Interval Computations. Netherlands: Kluwer Academic Publ., 1996. P. 337.
2. **Zadeh L. A.** Fuzzy sets // Information and Control. 1965. **8**, N 3. P. 338.
3. **Zadeh L. A.** Similarity relations and fuzzy orderings // Information Sciences. 1971.**3**. P. 177.
4. **Заде Л. А.** Понятие лингвистической переменной и его применение к принятию приближенных решений. М.: Мир, 1976.
5. **Zadeh L. A.** Theory of fuzzy sets // Encyclopedia of Computer Science and Technology /Eds. J. Belzer, A. Holzman. N. Y.: Marcel Decker. 1977.
6. **Zadeh L. A.** Fuzzy sets as a basis for a theory of possibility // Fuzzy Sets and Systems. 1978. **1**. P. 3.
7. **Кофман А.** Введение в теорию нечетких множеств: Пер. с фр. М.: Радио и связь, 1982.
8. **Dempster A. P.** Upper and lower probabilities induced by a multivalued mapping // Ann. Math. Statistics. 1967. **38**. P. 325.
9. **Shafer G.** A Mathematical Theory of Evidence. Princeton: Princeton Univ. Press, 1976.
10. **Dubois P., Prade H.** On several representations of an uncertain body of evidence // Fuzzy Information and Decision Processes /Eds. M. M. Gupta, E. Sanches. North-Holland, 1982. P. 167.
11. **Dubois P., Prade H.** Fuzzy sets and statistical data // European J. Operational Research. 1986. **25**. P. 345.
12. **Dubois P., Prade H.** A set-theoretic view of belief functions. Logical operations and approximations by fuzzy sets // Int. J. General Systems. 1986. **12(3)**. P. 193.
13. **Дюбуа Д., Прад А.** Теория возможностей. Приложения к представлению знаний в информатике: Пер. с фр. М.: Радио и связь, 1990.
14. **Плюта В.** Сравнительный многомерный анализ в эконометрическом моделировании: Пер. с польск. М.: Финансы и статистика, 1989.
15. **De Luca A., Termini S.** A definition of a non-probabilistic entropy in the setting of fuzzy sets theory // Information and Control. 1972. **20**. P. 301.

Поступила в редакцию 10 марта 1998 г.