УДК 311.2 + 616-036.22

А.В. Суворова

ПОДХОДЫ К ПРЕДСТАВЛЕНИЮ И ОБРАБОТКЕ НЕОПРЕДЕЛЕННОСТИ ДАННЫХ И ЗНАНИЙ О ПОВЕДЕНИИ ИНДИВИДОВ

Суворова А.В. **Подходы к представлению и обработке неопределенности данных и знаний о поведении индивидов.**

Аннотация. Предложен обзор средств представления и обработки неопределенности, которые могут оказаться полезными для решения задачи оценки интенсивности и производных характеристик поведения респондентов по их самоотчетам об эпизодах поведения. Рассмотрен вероятностный подход, байесовский подход, теория Демпстера–Шефера, теория нечетких множеств и их приложения к решению указанной задачи.

Ключевые слова: неопределенность, модели поведения, дефицит информации, вероятностный подход, теория Демпстера–Шефера, нечеткие множества, байесовский подход, неточность.

Suvorova A.V. **Approaches to the uncertainty representation and processing related to data about individual behavior.**

Abstract. We provide a description of the methods for representation and processing uncertainty that may be implemented to the problem of respondents' behavior rate estimate on the base of respondents' self-reports about last behavior episodes. We consider probability approach, Bayesian approach, Dempster–Shafer evidence theory, fuzzy sets theory and their application to the described problem.

Keywords: uncertainty, behavior models, information deficiency, probability approach, Dempster–Shafer evidence theory, fuzzy sets, Bayesian approach, imprecision.

1. Введение. Задачи моделирования социально-значимого поведения респондентов и разработки методов оценивания параметров такого поведения возникают во многих отраслях социологических, психологических, маркетинговых исследований [13, 16]. Это такие области, как лечение хронических заболеваний, эпидемиология ВИЧ/СПИД, оценка потребления тех или иных товаров или продуктов.

В частности, в современной эпидемиологии остро стоит вопрос об оценке риска передачи и приобретения опасных неизлечимых инфекций (например, ВИЧ) [16]. Наиболее точно такой риск характеризуется инсиденс-показателем (число заразившихся за определенный период среди лиц, подвергавшихся риску заражения, отнесенное к человеко-месяцам наблюдения). Для прямого измерения инсиденс-показателя требуется организовать когортное исследование, длящееся не менее полутора лет и подразумевающее вовлечение и сопровождение 500–1000 лиц из групп риска. Однократное проведение подобного когортного исследования обходится в полтора–два миллиона долларов. Такой уровень расходов делает невозможным мониторинг инсиденс-показателя даже в странах с сильной экономикой. Требуется предло-

жить математические модели, позволяющие выполнить более дешевые косвенные измерения инсиденс-показателя на основе ответов респондентов, составляющих выборку из группы риска [10, 11, 13, 14, 16].

При этом встает задача оценки интенсивности социально-значимого поведения респондента по его «одномоментному» самоотчету, то есть по ответам на блок вопросов или по результатам проведения интервью. Заметим, что подобные опросы опираются на информацию, хранящуюся в памяти респондента, и, естественно, чем глубже ретроспектива, тем труднее респондентам отвечать на вопросы и тем больше они делают ошибок припоминания.

Кроме того, полученные интервалы оказываются измеренными неточно, они характеризуются существенной недоопределенностью, связанной с тем, что результаты «измерения» зафиксированы на естественном языке, нечеткими терминами повседневной речи — с привычной и приемлемой для бытовых нужд гранулярностью, строгостью и точностью [10].

Следует подчеркнуть, что в силу неопределенности, неточности, нечеткости исходных данных, необходимости обрабатывать естественно-языковые «гранулярные» высказывания, «компенсировать» малочисленность статистических данных знаниями экспертов об особенностях того или иного вида социально-значимого поведения при решении всех подзадач станет необходимым использование моделей и методов для представления, обработки и комбинирования знаний с неопределенностью.

Цель статьи — рассмотреть те подходы к представлению и обработке неопределенности данных, сведений, знаний, которые могут оказаться полезными для формализации и последующего решения описанных выше проблем, возникающих при разработке процедур оценивания параметров социально-значимого поведения респондентов по их самоотчетам об эпизодах такого поведения.

2. Понятие неопределенности. Неопределенность — это в широком смысле отсутствие полной информации об интересующем объекте [48]. Использование понятия «неопределенность» во многих предметных областях приводит к тому, что единственного, подробного определения этого термина не существует. Так, например, в экономике есть такое определение: «неопределенность есть недостаточность сведений об условиях, в которых будет протекать экономическая деятельность» [8].

Вернемся к рассмотрению понятия неопределенности в информатике. Искусственный интеллект как научная дисциплина состоит из нескольких крупных течений. Одно из них — экспертные системы, представляющие собой направление исследований в области искус-

ственного интеллекта по созданию вычислительных систем, умеющих принимать решения, схожие с решениями экспертов в заданной предметной области [22]. Помимо собственно неполных, нечетких, неточных знаний, неопределенность может быть внесена и неточными или ненадежными данными о конкретной ситуации. Любой сенсор имеет ограниченную разрешающую способность и отнюдь не стопроцентную надежность. На практике далеко не всегда можно получить полные ответы на поставленные вопросы и, хотя можно воспользоваться различного рода дополнительной информацией о пациенте, например, с помощью дорогостоящих процедур или хирургическим путем, такие методики используются крайне редко из-за высокой стоимости и социальной значимости. Помимо всего прочего, существует еще и фактор времени. Таким образом, экспертным системам необходимо работать с недостаточной, неточной, неполной информацией [10, 11. 13].

В частности, в описанной выше задаче круглосуточный мониторинг числа эпизодов изучаемого поведения, как метод точного измерения, требует значительных временных, организационных ресурсов и в большинстве случаев невозможен. Косвенные же измерения не дают точных данных; кроме того, использование самоотчетов респондентов как наиболее доступного источника данных приводит к возникновению ошибок, связанных с необходимостью припоминания и с особенностями естественного языка, на котором формулируются ответы.

Природа неопределенности может быть классифицирована достаточно широко. Нижеприведенный подход классифицирует неопределенность в зависимости от информации и формы этой информации, которой располагает субъект при принятии решений [9]:

- неизвестность (незнание),
- физическая неопределенность,
- недостоверность (неполнота, недостаточность, не адекватность, расплывчатость),
- неоднозначность,
- лингвистическая неопределенность.

Иной подход к классификации неопределенностей приводится в книге Дж. фон Неймана и О. Моргенштерна «Теория игр и экономическое поведение» [7]:

- комбинаторное количество вариантов, которое просмотреть в отведенное время невозможно даже при наличии быстродействующих ЭВМ (полный перебор вариантов невозможен);
- случайные факторы происходящих событий как результат действия случайных сил: рассеяние попаданий в мишень при стрельбе; случайные потоки требований в обслуживающую

систему; случайные потоки денежных средств в банковскую систему или на предприятие и т.д.;

- стратегическая неопределенность (игровая неопределенность по существу) из-за неизвестного поведения противника (партнера — другого участника игры, включая игру с природой).

Еще несколько способов классификации неопределенностей указаны в [9]. В задачах управления различают неизвестность, неполноту и недостоверность информации, причем выделяются ситуации физической (объективной, стохастической) и лингвистической (субъективной, нестохастической) недостоверности.

В работах В.Б. Тарасова были отмечены (для инженерных знаний) [9]:

- неполнота, фрагментарность исходных данных и суждений, обусловленные недостаточным, неполным знанием характеристик новых материалов, особенностей технологических процессов и т.д.;
- нечеткость представлений о взаимосвязях между параметрами, связанная с отсутствием аналитических зависимостей между ними или вычислительной сложностью;
- ограничения по точности определения как количественных параметров, так и качественных факторов решения.

3. Вероятностный подход. Существуют различные подходы к представлению и обработке неопределенности данных и знаний, их описание можно найти во многих работах, в качестве примера можно привести [6, 15, 17, 26, 34, 40, 48, 50, 56]. Часть таких работ является историческими обзорами (например, в [17, 34, 48, 50, 56]), во многих был выполнен сравнительный анализ различных подходов, включающий как описание преимуществ, недостатков и областей применения каждого из подходов [33, 40, 56], так и возможность их применения для решения конкретной задачи, например [44, 49].

Исторически первым способом учета неопределенности были вероятности [34]. Существуют различные интерпретации вероятности [17]. Одна из них — частотная, вероятность рассматривается как относительная частота появления события, т.е. вероятность — объективное свойство ситуации. Такой подход используется, в частности в работах П. Лапласа, Р. Фишера, К. Пирсона, теории Неймана-Моргенштерна [7]. Субъективная интерпретация — вероятность как число, отражающее субъективное мнение или степень доверия к событию — была предложена Ф.Рамсеем [12, 17] и развита Л.Сэвиджем. В такой трактовке вероятность выражает познавательную активность исследователя или лица, вынужденного принимать решения в условиях дефицита информации. Существуют и другие способы интерпретации, например логический [28, 38].

Самые ранние работы учёных в области теории вероятностей относятся к XVII веку. Исследуя прогнозирование выигрыша в азартных играх, Б. Паскаль и П. Ферма открыли первые вероятностные закономерности, возникающие при бросании костей. Важный вклад в теорию вероятностей внёс Я. Бернулли: он дал доказательство закона больших чисел в простейшем случае независимых испытаний. В первой половине XIX века теория вероятностей начинает применяться к анализу ошибок наблюдений; П. Лаплас и С. Пуассон доказали первые предельные теоремы. Во второй половине XIX века основной вклад внесли русские учёные П.Л. Чебышев, А.А. Марков и А.М. Ляпунов. В это время были доказаны закон больших чисел, центральная предельная теорема, а также разработана теория цепей Маркова. Современный вид теория вероятностей получила благодаря аксиоматизации, предложенной А.Н. Колмогоровым. Более подробный исторический обзор можно посмотреть, например в [1, 2, 50].

Вероятностный подход к представлению и обработке неопределенности предполагает использование методов теории вероятности и математической статистики [1, 2, 18], то есть основан на понятии вероятности того, что некоторое событие произойдет, либо того, что некоторое логическое утверждение окажется истинным.

Заметим, что вопросами использования вероятности истинности для представления степени доверия к суждению (высказыванию) успешно занимался еще Дж. Буль [27], однако он не ставил, разумеется, вопросы автоматизации вероятностного и логико-вероятностного вывода и не рассматривал сложные модели на основе случайных бинарных либо многозначных элементов.

Представление неопределенности в виде скалярной оценки вероятности имеет ряд существенных несовершенств [17, 31, 34, 38, 50]. Среди них можно выделить следующие: точная вероятность неприемлема для полного отсутствия информации (ignorance); субъект может быть не готов определить вероятности для всех элементарных исходов, а только для некоторого их подмножества; использование вероятностей может потребовать значительных ресурсов. Одним из способов справиться с возникающими проблемами, оставаясь в рамках вероятностного подхода, является использование интервальных оценок вероятностей, например, вычисление верхней $\overline{P}(U)$ и нижней $\underline{P}(U)$ вероятностей [17, 34]

$$\overline{P}(U) = \sup\{p(U): p \in P\},$$
$$\underline{P}(U) = \inf\{p(U): p \in P\},$$

где P — множество распределений вероятностей на множестве Ω возможных исходов, $U \subseteq \Omega$.

Отметим, что интервальные оценки обладают рядом преимуществ по сравнению со скалярными. Они позволяют более полно описать неопределенность данных, кроме того, позволяют судить о точности полученных оценок (чем уже интервал, тем о большей точности мы можем говорить). Во многих задачах, возникающих на практике, нет необходимости в получении скалярной оценки, для принятия решения достаточно получить некоторый интервал возможных значений. Вместе с тем, применение интервальных оценок выливается в увеличение вычислительной сложности соответствующих алгоритмов вывода.

Одним из примеров возможного применения вероятностного подхода для решения задачи оценки параметров социально-значимого поведения по неполным, неточным и нечисловым данным является идея, изложенная в работах Н. В. Хованова[1]. Рассматривается конечное множество всех возможных значений некоторых показателей. На основе нечисловой (например, информации вида «значение показателя p_1 больше значения показателя p_2») и неточной (вида «значение показателя p_1 лежит в интервале (a_1, a_2)») информации строится множество всех допустимых значений показателей, т. е. значений, учитывающих эту информацию. Такая информация может не определять однозначно рассматриваемые значения, поэтому говорят о неполной информации.

Для моделирования неопределенности выбора конкретного значения из множества допустимых значений осуществляется рандомизация этого выбора: значение выбирается случайным образом, причем для каждого значения определена вероятностью такого выбора. Таким образом, набор показателей рассматривается как случайные величины, а в качестве оценки берется математическое ожидание этой величины [19, 20, 35].

Описанный метод в рамках задачи оценки интенсивности поведения можно проиллюстрировать следующим образом. Респонденты используют в своих высказываниях разные единицы измерения: часы, дни, недели, месяцы, полугода, года. Причем использованная единица измерения несет в себе информацию о точности измерения. Поясним это на примере двух, на первый взгляд равнозначных, высказываний: «семь дней назад» и «неделю назад». Когда респондент использует формулировку «семь дней назад», это свидетельствует о его уверенно-

[1] Н.В. Хованов решает задачи, в частности, связанные с формированием численных показателей, характеристик некоторой системы в условиях, когда недоопределен порядок на некотором множестве существенных для предметной области объектов, суждений или прогнозов.

сти в том, что событие произошло именно семь дней назад. В то время как «неделю назад» — это может быть и пять, и восемь дней назад [12, 15].

Для учета указанной неточности каждый ответ рассматривается не как точка на временной оси, а как интервал, длина которого зависит от единицы измерения. Значение каждого ответа рассматривается, таким образом, не как константа, а как случайная величина с заранее заданным распределением [13]. Введенная случайная величина за счет рандомизации [19] неопределенности ответа, обусловленной нечеткостью его формулировки, позволяет рассмотреть интенсивность как случайную величину и вычислить характеристики последней.

Одним из возможных обобщений классического вероятностного подхода можно считать модель неточных вероятностей (imprecise probabilities) [52].

4. Теория Демпстера–Шефера. Еще одним подходом [17, 34, 48] к получению интервальных вероятностей является теория Демпстера–Шефера — математическая теория очевидностей (свидетельств), основанная на мере доверия (belief functions) и мере правдоподобия (plausible reasoning), которые используются, чтобы скомбинировать отдельные части информации (свидетельства) для вычисления вероятности события. Теория была развита А. Демпстером (A. Dempster) и Г. Шефером (G. Shafer) [31, 34, 51]

Теория Демпстера–Шефера разработана для того, чтобы можно было учесть различие между неопределенностью (uncertainty) и незнанием (ignorance). Вместо вычисления вероятности высказывания в ней вычисляется вероятность того, что данное свидетельство поддерживает высказывание. Этот показатель измерения степени уверенности называется мерой доверия (belief function), которая обозначается как $\mathrm{Bel}(X)$ и изменяется от нуля, что означает отсутствие свидетельств в пользу X, до единицы, указывающей на определенность. Функция правдоподобия определяется следующим образом $\mathrm{Pl}(X)=1-\mathrm{Bel}(not(X))$ и также измеряется от нуля до единицы [6, 15, 17, 31, 34, 51].

При этом классическая вероятность $P(X)$ того, что X содержит истинную гипотезу, находится внутри интервала $[\mathrm{Bel}(A);\mathrm{Pl}(A)]$: $\mathrm{Bel}(X)\le P(X)\le \mathrm{Pl}(X)$. Если вероятности истинности гипотез полностью известны, эта теория сводится к классической теории вероятностей, при чем $\mathrm{Bel}(X)=P(X)=\mathrm{Pl}(X)$.

Следовательно, ширина доверительного интервала может служить оценкой неуверенности в справедливости гипотез из подмножества X

при имеющемся наборе свидетельств. Например, предположим, что в задаче существуют две взаимоисключающие гипотезы h_1 и h_2. Когда нет информации в пользу истинности какой-либо из гипотез, то обе они оцениваются одним и тем же интервалом [0, 1]. В процессе получения свидетельств истинности или ложности гипотез мы вправе ожидать, что интервалы будут сжиматься, выражая изменение степени нашей уверенности в значении истинности гипотез [6, 17, 34].

Пусть задано Q — множество взаимно исключающих гипотез (высказываний о состоянии мира), называемое фреймом различений (frame of discernment). Определим [17, 31] весовую функцию m следующим образом $m: 2^Q \to [0;1], \sum_{X \subset Q} m(X) = 1, \forall X \subset Q\ m(X) \geq 0$. Тогда мера доверия для $X \subset Q$ вводится как $\mathrm{Bel}(X) = \sum_{X_i \subset X} m(X_i)$ [6, 15, 17, 31, 34, 51].

Теория основана на двух идеях: получении степени доверия для данной задачи из субъективных свидетельств и использовании правила объединения свидетельств, предложенном в [31].

Таким образом, теория Демпстера–Шефера позволяет выразить отличие между частичной уверенностью и полным незнанием [6, 17, 34]. Одним из вариантов развития теории Демпстера–Шефера является теория возможностей и необходимостей [25, 32].

5. Байесовский подход. Другим методом обработки и представления неопределенности, так же, как и теория Демпстера–Шефера, использующим понятия свидетельств и гипотез, является байесовский подход.

Предполагается, что еще до получения свидетельств с каждой отдельной гипотезой связано определенное значение степени доверия к ее правдоподобности. При данном подходе степень доверия, назначенная множеству гипотез, распределяется между членами этого множества как функция их априорных вероятностей. В то же время степень доверия, назначенная группе гипотез, является суммой соответствующих показателей элементов этой группы. Обновление значений показателей доверия может выполняться рекурсивно, т.е. апостериорные оценки, полученные на основании одних свидетельств, могут использоваться в качестве априорных оценок для следующего цикла обновления при получении новых свидетельств. Как только данные получены, распределение вероятностей пересчитывается, учитывая новую информацию.

Основой данного подхода является теорема Байеса [1, 2, 18, 21].

Пусть мы наблюдаем случайную величину Y с плотностью вероятности $p(y \mid \theta)$, где θ — параметр такого распределения. Допустим, что выводы после наблюдения мы хотим сделать о величине θ, имеющей плотность распределения $\pi(\theta)$. Тогда, используя формулу Байеса, можно получить $\Pr(\theta \mid y) = \dfrac{\Pr(y \mid \theta)\Pr(\theta)}{\Pr(y)}$.

Обобщая на случай n переменных, получим

$$\Pr(\theta_j \mid y) = \frac{\Pr(y \mid \theta_j)\Pr(\theta_j)}{\Pr(y)} = \frac{\Pr(y \mid \theta_j)\Pr(\theta_j)}{\sum_{i=1}^{n} \Pr(y \mid \theta_i)\Pr(\theta_i)}.$$

Распределение $\Pr(\theta)$ называется априорным распределением θ, а $\Pr(\theta \mid y)$ — апостериорным распределением случайной величины θ при условии наблюдений y. Переменные θ_j называются гипотезами, а y — свидетельствами, поддерживающими гипотезы.

Аналогичные результаты можно получить и для непрерывного случая

$$\Pr(\theta \mid y) = \frac{p(y \mid \theta)\pi(\theta)}{p(y)} = \frac{p(y \mid \theta)\pi(\theta)}{\int p(y \mid \theta) d\theta},$$

где $\pi(\theta)$ — априорная плотность θ.

Наибольшее распространение получили три метода перехода от вероятностных распределений к точечным оценкам $\hat{\theta}$ параметра. К ним относятся [17]:

- мода $p\left(\hat{\theta} \mid y\right) = \max\limits_{\theta} p(\theta \mid y)$;
- медиана $\int\limits_{\hat{\theta}}^{+\infty} p(\theta \mid y) d\theta = \int\limits_{-\infty}^{\hat{\theta}} p(\theta \mid y) d\theta = 0{,}5$;
- среднее (математическое ожидание) $\hat{\theta} = \mathrm{E}(\theta \mid y)$.

Данный подход широко используется для представления неопределенности, оценки параметров, моделирования. В частности, на нем базируется математический аппарат байесовских сетей доверия, разработанный Дж. Перлом [45, 46]. С точки зрения цели настоящей работы, задачу об оценке интенсивности социально-значимого поведения [13, 14, 16] удобно свести через ряд промежуточных шагов к разработке особой вероятностной графической модели класса байесовских сетей доверия [14, 29, 30, 36, 39, 46]. При этом следует отметить, что исходная задача может не иметь аналитического решения, либо такое

решение будет весьма громоздким, либо потребует долгих поисков представления в "удобной и красивой" форме. Сведение же исходной задачи к построению байесовской сети доверия позволяет воспользоваться уже существующим мощным алгоритмическим аппаратом теории байесовских сетей доверия и свободно распространяемым программным инструментарием для проведения вычислительных экспериментов, а затем, и для использования построенной модели в практических целях.

6. Нечеткие множества. Еще один широко распространенный и быстроразвивающийся подход к представлению неполноты знаний в экспертных системах — это нечеткость. Многими исследователями выдвигались аргументы [5, 6, 15, 17, 42, 50], показывающие, что теория вероятности не является адекватным инструментом для решения задач представления неопределенности знаний и данных. Перечислим некоторые из них:

- теория вероятности не дает ответа на вопрос, как комбинировать вероятности с количественными данными;
- назначение вероятности определенным событиям требует информации, которой мы просто не располагаем;
- непонятно, как количественно оценивать такие часто встречающиеся на практике понятия, как "в большинстве случаев", "в редких случаях", или такие приблизительные оценки, как "старый" или "высокий";
- применение теории вероятности вынуждает давать точные оценки тем параметрам, которые они не могут оценить;
- обновление вероятностных оценок обходится очень дорого, поскольку требует большого объема вычислений.

Все эти соображения породили новый формальный аппарат для работы с неопределенностями, который получил название нечеткая логика (fuzzy logic).

Теория нечетких множеств была заложена в фундаментальных работах Лофти Заде [3, 4, 54, 55] Первоначальным замыслом этой теории было построить функциональное соответствие между нечеткими лингвистическими описаниями (типа "высокий", "теплый" и т.д.) и специальными функциями, выражающими степень принадлежности значений измеряемых параметров (длины, температуры, веса и т.д.) упомянутым нечетким описаниям. В [3], были введены так называемые лингвистические вероятности — вероятности, заданные не количественно, а при помощи нечетко-смысловой оценки.

Лингвистическая переменная отличается от числовой переменной тем, что ее значениями являются не числа, а слова или предложения в естественном или формальном языке. Поскольку слова в общем менее

точны, чем числа, понятие лингвистической переменной дает возможность приближенно описывать явления, которые настолько сложны, что не поддаются описанию в общепринятых количественных терминах. В частности, нечеткое множество, которое представляет собой ограничение, связанное со значениями лингвистической переменной, можно рассматривать как совокупную характеристику различных подклассов элементов универсального множества. В этом смысле роль нечетких множеств аналогична той роли, которую играют слова и предложения в естественном языке. Например, прилагательное "*красивый*" отражает комплекс характеристик внешности индивидуума. Это прилагательное можно также рассматривать как название нечеткого множества, которое является ограничением, обусловленным нечеткой переменной «*красивый*». С этой точки зрения термины «*очень красивый*», «*некрасивый*», «*черезвычайно красивый*», «*вполне красивый*» и т.п. — названия нечетких множеств, образованных путем действия модификаторов «*очень, не, черезвычайно, вполне*» и т.п. на нечеткое множество «*красивый*». В сущности, эти нечеткие множества вместе с нечетким множеством «*красивый*» играют роль значений лингвистической переменной «*внешность*» [3, 5].

Лингвистическая переменная характеризуется набором свойств $\langle X, T(X), U, G, M \rangle$, в котором X — название переменной; $T(X)$ обозначает терм-множество переменной X, т.е. множество названий лингвистических значений переменной X, причем каждое из таких значений является нечеткой переменной x со значениями из универсального множества U с базовой переменной u; G — синтаксическое правило, порождающее названия х значений переменной X; M — семантическое правило, которое ставит в соответствие каждой нечеткой переменной x ее смысл $M(x)$, т.е. нечеткое подмножество $M(x)$ универсального множества U [3, 5, 42].

В экспертных системах, основанных на нечеткой логике, поведение исследуемой системы описывается в естественном (или близком к естественному) языке в терминах лингвистических переменных. Входные и выходные параметры системы рассматриваются как лингвистические переменные.

Преимущество данной модели — в ее универсальности. Нам неважно, что именно на входе — конкретные числовые значения или некоторая неопределенность, описываемая нечетким множеством. Но за данную универсальность приходится расплачиваться сложностью системы. Поэтому этой общей моделью на практике пользуются довольно редко. Обычно же используют ее упрощенный вариант, называемый нечетким выводом. Он основывается на предположении, что

все входные лингвистические переменные имеют известные нам числовые значения (как и бывает довольно часто на практике). Также обычно не используют более одной выходной лингвистической переменной [42].

Нечетким логическим выводом (fuzzy logic inference) называется аппроксимация зависимости $Y = f(X_1, X_2, ..., X_n)$ каждой выходной лингвистической переменной от входных лингвистических переменных и получение заключения в виде нечеткого множества, соответствующего текущим значениям входов, с использованием нечеткой базы знаний и нечетких операций, описанных в [54].

Диапазон применимости теории нечетких множеств существенно расширился. Сам Л. Заде определил нечеткие множества как инструмент построения теории возможностей [32, 55]. Следующим достижением теории нечетких множеств является введение в обиход так называемых нечетких чисел как нечетких подмножеств специализированного вида, соответствующих высказываниям типа "значение переменной примерно равно а". С их введением оказалось возможным прогнозировать будущие значения параметров, которые ожидаемо меняются в установленном расчетном диапазоне. Вводится набор операций над нечеткими числами, которые сводятся к алгебраическим операциям с обычными числами при задании определенного интервала достоверности (уровня принадлежности).

Области применения нечеткой логики заметно расширяются. Она применяется в автомобильной, аэрокосмической и транспортной промышленности, в области изделий бытовой техники, в сфере финансов, анализа и принятия управленческих решений и многих других [4, 5, 22-24, 40-42, 44, 47, 49, 50, 54, 55].

При решении задачи оценки интенсивности поведения респондентов по сведениям, полученным из их самоотчетов, мы имеем дело с данными, выраженными на естественном языке. Если естественно-языковую неопределенность ответов в форме «столько-то дней назад» достаточно легко представить и другими рассмотренными подходами к обработке неопределенности, в частности, вероятностным, то встречающиеся ответы «на прошлой неделе», «по праздникам» целесообразно обрабатывать именно средствами теории нечетких множеств.

7. Заключение. В данной работе предложен обзор средств представления и обработки неопределенности, которые могут оказаться полезными для решения задачи оценки интенсивности и производных характеристик поведения респондентов по их самоотчетам об эпизодах поведения. Рассмотрены вероятностный подход, байесовский подход, теория Демпстера–Шефера, теория нечетких множеств и их приложения к решению указанной задачи. Хотя в статье представлены не все

существующие методы, а лишь наиболее широко известные, они имеют непосредственное отношение к решению рассматриваемой задачи. Кроме перечисленных методов можно упомянуть, например, теорию неточных или приближенных множеств (rough sets) [43, 53], автор которой позиционирует ее как альтернативу теории нечетких множеств. Отметим, что предпринимаются попытки как обобщения классических подходов, и построения расширенной модели, например, [23, 24, 37], так и объединения достоинств нескольких подход в гибридные модели [25, 37].

Литература

1. *Боровков А.А.* Теория вероятностей. М.: Наука, 1976. 352 с.
2. *Вентцель Е.С.* Теория вероятностей. 2-е изд. М.: Наука, 1969. 576 с.
3. *Заде Л.* Понятие лингвистической переменной и ее применение к принятию приближенных решений. М.: Мир, 1976.
4. *Заде Л.* Роль мягких вычислений и нечеткой логики в понимании, конструировании и развитии информационных/интеллектуальных систем // Новости искусственного интеллекта. 2001. № 2–3(44–45). С. 7–11.
5. *Кофман А.* Введение в теорию нечетких множеств. М.: Радио и связь, 1982. 432 с.
6. *Люгер Дж.* Искусственный интеллект: стратегии и методы решения сложных проблем. 4-е изд. М.: Издательский дом "Вильямс", 2005. 864 с.
7. *Нейман Дж., Моргенштерн О.* Теория игр и экономическое поведение. М.: Наука, 1970. 707 с.
8. Неопределенность в экономике. 14 определений понятия неопределенность. Неопределенность ее виды. - Web: http://www.beintrend.ru/menedzhment/
9. Нечеткие гибридные системы. Теория и практика / под ред. Ярушкиной Н.Г. - М.: Физматлит, 2007. 208 с.
10. Пащенко А. Е., Тулупьев А. Л., Тулупьева Т. В. Базисная темпоральная онтология для обработки ответов об участии в рискованном поведении, связанном с передачей ВИЧ // Научная сессия МИФИ-2008. Сб. научн. трудов. В 15 томах. Т. 10. Интеллектуальные системы и технологии. М.: МИФИ, 2008. С. 109–111.
11. *Пащенко А.Е., Тулупьев А.Л., Тулупьева Т.В., Красносельских Т.В., Соколовский Е.В.* Косвенная оценка вероятности заражения ВИЧ-инфекцией на основе данных о последних эпизодах рискованного поведения // Здравоохранение Российской Федерации. 2010. № 2. С. 32–35.
12. Субъективная вероятность (обзор). Web: http://ru.science.wikia.com.
13. *Суворова А.В., Тулупьев А.Л., Пащенко А.Е., Тулупьева Т.В., Красносельских Т.В.* Анализ гранулярных данных и знаний в задачах исследования социально значимых видов поведения // Компьютерные инструменты в образовании. №4. 2010. С. 30–38.
14. *Суворова А.В., Тулупьева Т.В., Тулупьев А.Л., Сироткин А.В., Пащенко А.Е.* Вероятностные графические модели социально-значимого поведения индивида, учитывающие неполноту информации // Труды СПИИРАН. 2012. Вып. 3 (22). С. 101–112.
15. *Тулупьев А.Л., Николенко С.И., Сироткин А.В.* Байесовские сети: логико-вероятностный подход. СПб.: Наука, 2006. 607 с.
16. *Тулупьева Т.В., Пащенко А.Е., Тулупьев А.Л., Красносельских Т.В., Казакова О.С.* Модели ВИЧ-социально-значимого поведения в контексте психологической защиты и других адаптивных стилей. СПб.: Наука, 2008. 140 с.
17. *Уткин Л.В.* Анализ риска и принятие решений при неполной информации. СПб.: Наука, 2007. 404 с.
18. *Феллер В.* Введение в теорию вероятностей и ее приложения. М.: Мир, 1984. 730 с.

19. *Хованов Н.В.* Анализ и синтез показателей при информационном дефиците. СПб.: Изд-во СПбГУ, 1996. 196 с.
20. *Хованов Н.В.* Метод рандомизированных траекторий в задачах оценки функциональной зависимости // Труды СПИИРАН. 2009. Вып. 9. СПб.: Наука, 2009. С. 262–279.
21. *Ширяев А.Н.* Вероятность: Учеб.пособ. для вузов. 2-е изд. М.: Наука. Гл. ред. физ.-мат. лит., 1989. 640 с.
22. Экспертные системы. Web: http://www.aiportal.ru/articles/expert-systems/expert-systems.html.
23. *Atanassov K.* Intuitionistic Fuzzy Sets // Fuzzy Sets and Systems. 1986. Vol. 20. Pp.87-96.
24. *Atanassov K.* More on Intuitionistic Fuzzy Sets // Fuzzy Sets and Systems. 1989. Vol. 33. Pp.37-45.
25. *Baudrit C., Cousco I., Dubois D.* Joint propagation of propability and possibility in risk analysis: Towards a formal framework // International Journal of Approxomate Reasoning. 2007. Vol. 45. Pp.82-105.
26. *Booker J., Ross T.* An evolution of uncertainty assessment and quantification // Scientia Iranica. 2011. Vol. 18. No. 3. Pp. 669-676.
27. *Bool G.* An Investigation of the Laws of Thought, on Which Are Founded the Mathematical Theories of Logic and Probabilities. Cambridge: Macmillan / London: Walton &Maberly, 1854. (Reprinted in 1951, Dover Publications, New York.)
28. *Carnap R.* Logical foundations of probability. Chicago: Chicago university Press, 1950. 607 p.
29. *Cowell R.G., Dawid A.Ph., Lauritzen S.L., Spiegelhalter D.J.* Probabilistic Networks and Expert Systems. Berlin: Springer, 2003. 321 p.
30. *Darwiche F.* Modeling and reasoning with Bayesian networks. Cambridge: Cambridge University Press, 2009. 548 c.
31. *Dempster D.* Upper and lower probabilities induced by a multi-valued mapping // Annals of Mathematical Stattistics, 1967. Vol.38. Pp. 325-339.
32. *Dubois D., Prade H.* Possibility Theory: An Approach to Computerized Processing of Uncertainty. NY: Plenum Press, 1988
33. *Fine T.L.* Lower probability models for uncertainty and non-deterministic processes // Journal of statistical Planning and inference. 1988. Vol. 20. P. 389-411
34. *Halpern J.* Reasoning about Uncertainty. Massachusetts: The MIT Press, 2003. 497 p.
35. *Hovanov N., Yudaeva M., Hovanov K.* Multicriteria estimation of probabilities on basis of expert non-numeric, non-exact and non-complete knowledge // European Journal of Operational Research. 2009. № 195. P. 857–863.
36. *Jensen F.V.* Bayesian Networks and Decision Graphs. NY.: Springer-Verlag, 2001. 268 p.
37. *Joslyn C., Rocha L.* Towards a formal taxonomy of hybrid uncertainty representations // Information Sciences. 1998. Vol. 110. Pp. 255-277.
38. *Keynes J.M.* A treatise on probability. London: Macmillan, 1921. 459 p.
39. *Korb K.B., Nicholson A.E.* Bayesian Artificial Intelligence. NY.: Chapman and Hall/CRC, 2004. 364 p.
40. *Laskey K., Levitt T.* Artificial Intelligence: Uncertainty // International Encyclopedia of the Social & Behavioral Sciences, 2001. Pp. 799-805
41. *Nguyen H.* Some mathematical tools for linguistic probabilities // Fuzzy Sets and Systems. 1979. Vol. 2. Pp.53-65.
42. *Nguyen H.T., Walker E.A.* A First Course in Fuzzy Logic. Second ed. NY etc.: Chapman&Hall/CRC, 2000. 373 p.
43. *Pawlak Z.* Rough Sets // International Journal of Computer and Information Sciences. 1982. Vol. 11. No. 5. Pp. 341-456.

44. *Pei Z., Zheng L.* A novel approach to multi-attribute decision making based on intuitionistic fuzzy sets // Expert Systems with Applications. 2012. Vol. 39. Pp.2560-2566.
45. *Perl J.* Causality: Models, Reasoning, and Inference. Cambridge: Cambridge University Press, 2000. 400 c.
46. *Perl J.* Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. NY etc.: Morgan Kaufmann Publ., 1994. P. 552.
47. *Roldan C., Roldan A., Martinez-Moreno J.* A fuzzy regression model based on distances and random variables with crisp input and fuzzy output data: a case study in biomass production // Soft Computing. 2012. Vol. 16. Pp. 785-795.
48. *Ross T., Booker J., Montoya A.* New developments in uncertainty assessment and uncertainty management // Expert Systems with Applications. 2013. Vol. 40. Pp. 964-974.
49. *Rossi R., Gastaldi M., Gecchele G., Meneguzzer C.* Comparative analysis of random utility models and fuzzy logic models for representing gap-acceptance behavior using data from driving simulator experiments // Procedia — Social and Behavioral Sciences. 2012. Vol. 54. Pp. 834-844.
50. *Rouvray D.* The treatment of uncertainty in the sciences // Endeavour. 1997. Vol. 21. No. 4. Pp. 154-158.
51. *Shafer G.* A Mathematical Theory of Evidence. Princeton, N.J.: Princeton University Press, 1976. 297 p.
52. *Walley P.* Statistical reasoning with imprecise probabilities. NY, London: Chapman and Hall, 1991. xii + 709 p.
53. *Wei W., Liang J., Qian Y.* A comparative study of rough sets for hybrid data // Information Sciences. 2012. Vol. 190. Pp. 1-16.
54. *Zadeh L.A.* Fuzzy Sets // Information and Control. 1965. Vol. 8. Pp. 338-353.
55. *Zadeh L.A.* Fuzzy sets as a basis for a theory of possibility // Fuzzy Sets and Systems. 1978. Vol.1. №1.
56. *Zhang Q., Xiao Y., Xing Y.* The representation and processing of uncertain problems //Procedia Engineering. 2011. Vol. 15. Pp. 1958-1962.

Суворова Алена Владимировна — младший научный сотрудник лаборатории теоретических и междисциплинарных проблем информатики, СПИИРАН, аспирант математико-механического факультета СПбГУ. Область научных интересов: математическая статистика, теория вероятности, применение методов математического моделирования в эпидемиологии. Число научных публикаций — 21. SuvorovaAV@iias.spb.su, www.tulupyev.spb.ru; СПИИРАН, 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178, РФ; р.т. +7(812)328-3337, факс +7(812)328-4450. Научный руководитель — А.Л. Тулупьев.

Suvorova Alena Vladimirovna — junior researcher, Laboratory of Theoretical and Interdisciplinary Computer Science, SPIIRAS, PhD student, Faculty of Mathematics and Mechanics of SPbSU. Research interests: mathematical statistics, probability theory, application of mathematical modeling in epidemiology. The number of publications — 21. SuvorovaAV@iias.spb.su, www.tulupyev.spb.ru; SPIIRAS, 39, 14th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)328-3337, fax +7(812)328-4450. Scientific advisor — A.L. Tulupiev.

Поддержка исследований. Статья содержит материалы исследований, частично поддержанных грантами РФФИ 09-01-00861-а, 10-01-00640-а, 12-01-00945-а, грантом для молодых ученых и кандидатов наук Комитета по науке и высшей школе Правительства Санкт-Петербурга «Модели и алгоритмы анализа сверхкоротких неточных временных рядов на основе гранулярных данных и знаний», руководитель — А.В. Суворова (2012).

Рекомендовано ТиМПИ СПИИРАН, зав. лаб. д-р физ.-мат. наук, доцент А.Л. Тулупьев.
Статья поступила в редакцию 23.09.2012.

РЕФЕРАТ

Суворова А.В. **Подходы к представлению и обработке неопределенности данных и знаний о поведении индивидов.**

Задачи моделирования социально-значимого поведения респондентов и разработки методов оценивания параметров такого поведения возникают во многих отраслях социологических, психологических, маркетинговых исследований. При этом встает задача оценки интенсивности социально-значимого поведения респондента по его «одномоментному» самоотчету, то есть по ответам на блок вопросов или по результатам проведения интервью. Заметим, что подобные опросы опираются на информацию, хранящуюся в памяти респондента, и, естественно, чем глубже ретроспектива, тем труднее респондентам отвечать на вопросы и тем больше они делают ошибок припоминания.

Полученные интервалы измерены неточно, они характеризуются существенной недоопределенностью, связанной с тем, что результаты «измерения» зафиксированы на естественном языке, нечеткими терминами повседневной речи — с привычной и приемлемой для бытовых нужд гранулярностью, строгостью и точностью.

Следует подчеркнуть, что в силу неопределенности, неточности, нечеткости исходных данных, необходимости обрабатывать естественно-языковые "гранулярные" высказывания, "компенсировать" малочисленность статистических данных знаниями экспертов об особенностях того или иного вида социально-значимого поведения при решении всех подзадач станет необходимым использование моделей и методов для представления, обработки и комбинирования знаний с неопределенностью.

В работе предложен обзор средств представления и обработки неопределенности, которые могут оказаться полезными для решения задачи оценки интенсивности и производных характеристик поведения респондентов по их самоотчетам об эпизодах поведения. Рассмотрен наиболее широко известные методы: вероятностный подход, байесовский подход, теория Демпстера–Шефера, теория нечетких множеств и их приложения к решению указанной задачи.

SUMMARY

Suvorova A.V. **Approaches to the uncertainty representation and processing related to data about individual behavior.**

In many fields of sociological, psychological and marketing research, we face the problem of socially significant behavior rate or frequency estimate. We need to estimate behavior rate on the base of the respondent's «one-stage» self-reporting, or, other words, on the base of the responses to the questionnaire or the results of the interview. Note that these sets of questions are based on information stored in the respondent's memory, and, of course, if the retrospective is deeper, when it is more difficult for the respondents to answer questions and it is more likely they make errors.

The results of the respondents' interview about their behavior come in everyday language, that's why this data is fuzzy and incomplete. Answers about the last episodes have a special characteristic: they are granular, because in our everyday activity we use not exact time estimates. Such data statements should be organized, categorized and formalized for further processing. Limited number of natural language responses' forms and their fuzziness, uncertainty, imprecision do not allow direct use of known methods of queuing theory for behavior rate estimate.

We provide a description of the methods for representation and processing uncertainty that may be implemented to the problem of respondents' behavior rate estimate on the base of respondents' self-reports about last behavior episodes. We consider probability approach, Bayesian approach, Dempster–Shafer evidence theory, fuzzy sets theory and their application to the described problem.