О СВОЙСТВАХ РАССТОЯНИЙ В ВЕРОЯТНОСТНЫХ ВЫСКАЗЫВАНИЯХ ЭКСПЕРТОВ *

Александр Викентьев, Руслан Викентьев

Abstract: The paper discusses probabilities' logical expert statements represented as the formulas of Sentence Logic. Methods for setting metrics on such formulas are offered and the entered metric's properties are investigated. The research can be applied to solving the problems of the best reconciliation of expert statements, to constructing the decision functions in pattern recognition and building the expert systems.

Keywords: *cluster analysis, expert statements, distance, metrics*

ACM Classification Keywords: *I.2.6. Artificial Intelligence - knowledge acquisition.*

Введение

Поскольку сейчас проявляется все больший интерес к анализу экспертной информации, заданной в виде вероятностных логических высказываний экспертов, интересны также вопросы о высказываниях экспертов, представленных формулами исчисления высказываний с вероятностями. Возникают задачи об алгоритмах распознавания закономерностей, согласования таких знаний и их кластеризации [Блощицын В.Я.., Лбов, 1987 - Викентьев, Лбов, 1998]. Для этого необходимы метрики на знаниях. В работе рассматриваются логические высказывания экспертов, представленные формулами исчисления высказываний (ИВ) с вероятностями.

Вопросами введения расстояний на высказываниях экспертов и их применениями занимался профессор Г.С. Лбов [Лбов, Старцева,1999], использовал для построения решающих функций и для согласования высказываний совместно со своими учениками, использует в своих работах д.т.н. В.Б. Бериков [Бериков, 2009 -- Лбов, Бериков, 2005], разрабатывал и продолжает применять в алгоритмах распознавания профессор Н.Г. Загоруйко и один из авторов статьи.

В данной статье предлагаются способы задания расстояний на логических высказываниях с вероятностями с использованием логических моделей, вообще говоря, неполной теории. Если класс моделей теории эффективно задан, то мы можем все нужные расстояния найти, а также вычислить вероятности выполнимости любой формулы на данном конечном классе. Нами здесь рассматривается обратная задача: мы знаем от экспертов вероятности формул-высказываний, знаем как-то логическую теорию, которая задает модели нашей исследуемой области и хотим восстановить расстояния между данными формулами, мы его не знаем и хотим узнать. Ясно, что такая задача может иметь несколько решений из-за нехватки информации и про теорию и про адекватность ее моделей. В случае когда теория одна и модели ее устраивают всех экспертов, то решение будет единстственно.

Изучаются свойства введенных искомых расстояний. Для решения поставленной задачи используются вероятностный и теоретико - модельный подходы [Лбов, Старцева,1999 -- Vikent'ev, Lbov, 1997]. Результаты неоднократно доложены на международных конференциях в 2009, 2010 гг., в том числе юбилейной конференции к 100 - летию Академика Анатолия Ивановича Мальцева – выдающегося математика -- Логика и Алгебраиста, крупного ученого по алгебре, теории моделей, теории алгоритмов, теории нумерации, алгоритмическим проблемам и применениям методов математической логики.

.

* Работа выполнена при поддержке гранта РФФИ 10-01-00113а

РАССТОЯНИЯ НА ФОРМУЛАХ ИВ С ВЕРОЯТНОСТЯМИ И ИХ СВОЙСТВА

Будем рассматривать знания экспертов, представленные формулами ИВ с вероятностями (вероятностные высказывания), т.е. высказывания вида: "ϕ с вероятностью p_ϕ", где ϕ- формула ИВ. Используем запись для таких высказываний: $B_i = \langle \phi, p_\phi \rangle$, $B_j = \langle \psi, p_\psi \rangle$.

Знакомство с математической логикой предполагается [Ершов, Палютин, 2004]

Пусть Σ - база знаний, состоящая из формул ИВ (в Σ содержатся все формулы, с которыми будут работать эксперты, а также конечное число аксиом теории). $S(\phi)$ - носитель формулы ϕ, т.е. множество элементарных высказываний, используемых при написании формулы ϕ. $S(\Sigma) = \bigcup_{\phi \in \Sigma} S(\phi)$ - носитель совокупности знаний. Для простототы будем считать, что он конечен.

Рассмотрим множество $P(S(\Sigma)) = 2^{S(\Sigma)}$- множество всевозможных подмножеств множества $S(\Sigma)$. Элементы множества $P(S(\Sigma))$ назовем моделями. Известно, что $|P(S(\Sigma))| = 2^{|S(\Sigma)|} = n$ (-- для простоты обозначения). Модели теории образуют подмножество этого множества и для его мощности, в дальнейшем, будем использовать тоже ***n***.

Пусть эксперты говорят о вероятностях (частости) формул на множестве n моделей, и каждое высказывание присутствует только с одной вероятностью.

Тогда будем интерпретировать вероятность, данную экспертом, следующим образом: $B = \langle \phi, p_\phi \rangle$ означает, что высказывание ϕ истинно на $n_\phi = [n \cdot p_\phi]$ моделях, где n - число моделей рассматриваемой теории..

Пусть у нас есть два вероятностных логических высказывания $B_i = \langle \phi, p_\phi \rangle$ и $B_j = \langle \psi, p_\psi \rangle$.

Зададим способ вычисления расстояния $\rho(B_i, B_j)$ между такими высказываниями.

Интерпретируя данные экспертами вероятности описанным выше способом, получаем, что высказывание ϕ истинно на $n_\phi = [n \cdot p_\phi]$ моделях, высказывание ψ истинно на $n_\psi = [n \cdot p_\psi]$ моделях. Предполагаем, что с этим подходом (аксиомами теории и интерпретацией вероятностей) согласны сами эксперты.

Отметим также, что (при таком подходе) мы точно не знаем на каких именно моделях каждое высказывание истинно, а также не знаем число моделей, на которых эти высказывания-пары истинны одновременно.

Будем решать такую задачу: пусть высказывание ϕ истинно на n_ϕ моделях, высказывание ψ истинно на n_ψ моделях и k - число моделей, на которых эти высказывания истинны одновременно. Тогда как находить и вычислять расстояние между высказываниями $B_i = \langle \phi, p_\phi \rangle$ и $B_j = \langle \psi, p_\psi \rangle$?

Обозначим возникающие далее расстояния за $\rho_k(B_i, B_j)$, где $k = 0,1,2,\ldots,\min(n_\phi, n_\psi)$.

Как и раньше [Лбов, Старцева,1999 -- Викентьев, Лбов,1998], расстояние $\rho_k(B_i, B_j)$ определим через симметрическую разность моделей на которых они истинны, то есть

$\rho_k(B_i, B_j) = \dfrac{n_\phi - k + n_\psi - k}{n} = \dfrac{n_\phi + n_\psi - 2k}{n}$, для каждого $k = 0,1,2,\ldots,\min(n_\phi, n_\psi)$.

Теорема 1. Для расстояний $\rho_k(B_i, B_j)$ справедливы свойства:

$0 \le \rho_k(B_i, B_j) \le 1$.

$\rho_k(B_i, B_j) = \rho_k(B_j, B_i)$.

$\rho_k(B_i, B_j) \le \rho_k(B_i, B_s) + \rho_k(B_s, B_j)$.

$B_i \equiv B_j \Leftrightarrow \rho_k(B_i, B_j) = 0$ ($B_i \equiv B_j \Leftrightarrow \phi \equiv \psi$ и $p_\phi = p_\psi$ - это означает, что формулы ϕ и ψ истинны на одних и тех же моделях).

$B_i \equiv \neg B_j \Leftrightarrow \rho_k(B_i, B_j) = 1$.

$\rho_k(B_i, B_j) = 1 - \rho_k(B_i, \neg B_j) = \rho_k(\neg B_i, \neg B_j)$.

$\rho_k(B_i, B_j) = \rho_k(B_i \wedge B_j, B_i \vee B_j)$.

Докажем неочевидное свойство 3. Определение $\rho_k(B_i, B_j)$ можно переписать следующим образом:

$\rho_k(B_i, B_j) = \dfrac{n_\phi + n_\psi - 2k}{n} = \dfrac{n_{\phi\Delta\psi}}{n} = \dfrac{n_{(\neg\phi\wedge\psi)\vee(\phi\wedge\neg\psi)}}{n}$ (по определению симметрической разности). Тогда для произвольного высказывания $B_s = \left\langle \chi, p_\chi \right\rangle$ нетрудно доказывается, что $n_{\phi\Delta\psi} \le n_{\phi\Delta\chi} + n_{\chi\Delta\psi}$. Тогда

$$\rho_k(B_i, B_j) = \frac{n_{\phi\Delta\psi}}{n} \le \frac{n_{\phi\Delta\chi}}{n} + \frac{n_{\chi\Delta\psi}}{n} = \rho_k(B_i, B_s) + \rho_k(B_s, B_j).$$

Докажем так же свойство 4. Докажем сначала слева направо ($\Longrightarrow$). Если $B_i \equiv B_j$, то $\phi \equiv \psi$ и, значит, $n_\phi = n_\psi = k$. Следовательно, $\rho_k(B_i, B_j) = \dfrac{n_\phi + n_\psi - 2k}{n} = 0$.

Докажем теперь в обратную сторону($\Longleftarrow$). Если $\rho_k(B_i, B_j) = 0$, то $n_\phi + n_\psi - 2k = 0$. Так как k может принимать значения $0,1,2,\ldots,\min(n_\phi, n_\psi)$, то $(n_\phi + n_\psi - 2k = 0 \Leftrightarrow n_\phi = n_\psi = k)$. Следовательно, $\phi \equiv \psi$ и значит $B_i \equiv B_j$. Докажем свойство 5. Докажем ($\Longrightarrow$). Если $B_i \equiv \neg B_j$, то $\phi \equiv \neg\psi$. Тогда $n_\phi = n - n_\psi$ и $k = 0$, следовательно, $\rho_k(B_i, B_j) = \dfrac{n_\phi + n_\psi - 0}{n} = \dfrac{n}{n} = 1$.

Докажем ($\Longleftarrow$). Если $\rho_k(B_i, B_j) = 1$, то $n_\phi + n_\psi - 2k = n$. Так как k может принимать значения $0,1,2,\ldots,\min(n_\phi, n_\psi)$, то $(n_\phi + n_\psi - 2k = n \Leftrightarrow n_\phi + n_\psi = n$ и $k = 0)$. Следовательно, $\phi \equiv \neg\psi$ и $B_i \equiv \neg B_j$.

Докажем свойство 6. $\rho_k(B_i, B_j) = \dfrac{n_\phi + n_\psi - 2k}{n}$. Рассмотрим

$$\rho_k(B_i, \neg B_j) = \frac{n_\phi + n_{\neg\psi} - 2(n_\phi - k)}{n} = \frac{n_\phi + (n - n_\psi) - 2(n_\phi - k)}{n} =$$

$$= \frac{n - n_\phi - n_\psi - 2k}{n} = 1 - \frac{n_\phi + n_\psi - 2k}{n} = 1 - \rho_k(B_i, B_j).$$

Доказывается нетрудно, что $1 - \rho_k(B_i, \neg B_j) = \rho_k(\neg B_i, \neg B_j)$.

Докажем свойство 7. Можно доказать, что $n_{(\phi\wedge\psi)\Delta(\phi\vee\psi)} = n_{\phi\Delta\psi}$. Тогда

$$\rho_k(B_i \wedge B_j, B_i \vee B_j) = \frac{n_{(\phi\wedge\psi)\Delta(\phi\vee\psi)}}{n} = \frac{n_{\phi\Delta\psi)}}{n} = \rho_k(B_i, B_j).$$

Далее предложим несколько способов вычисления расстояния $\rho(B_i, B_j)$ между вероятностными высказываниями $B_i = \langle \phi, p_\phi \rangle$ и $B_j = \langle \psi, p_\psi \rangle$.

Так как нам не известно число k (число моделей, на которых высказывания ϕ и ψ истинны одновременно), и если нет никаких предпочтений для значения k (хотя оно и может быть высказано экспертами), то можем, например, поступить следующим образом.

Предположим так же, что для нас все значения для числа k равновероятны. Тогда расстояние между вероятностными высказываниями $B_i = \langle \phi, p_\phi \rangle$ и $B_j = \langle \psi, p_\psi \rangle$ определим как усреднение расстояний $\rho_k(B_i, B_j)$ по всем значениям k, т.е. $\rho(B_i, B_j) = \dfrac{\sum_{k=0}^{\min(n_\phi, n_\psi)} \rho_k(B_i, B_j)}{\min(n_\phi, n_\psi) + 1}$.

Для этого расстояния также справедлива теорема 1, и под знаком суммы слагаемые можно взять с весами с учетом общих мнений экспертов или адаптацией по задаче.

Если экспертами высказано, какое значение для k предпочтительнее, то в качестве $\rho(B_i, B_j)$ берем расстояние $\rho_k(B_i, B_j)$. Это так когда мы знаем, что пересечение состоит из k моделей.

Можно подойти к этому вопросу и с вероятностной точки зрения, построить статистико-вероятностную модель для каждого k и вычислить вероятность того, что высказывания ϕ и ψ одновременно истинны на k моделях.

Далее найдем вероятность p_k того, что в выбранных n_ϕ моделях и n_ψ моделях (они выбираются из n моделей) будет k моделей, на которых высказывания ϕ и ψ истинны одновременно, где $k = 0, 1, 2, \ldots, \min(n_\phi, n_\psi)$.

Сначала определим вероятностное пространство $\langle \Omega, \mathrm{A}, p \rangle$, где Ω- пространство элементарных исходов – моделей, $\Omega 1 = \{0, 1, 2, \ldots, \min(n_\phi, n_\psi)\}$ - число возможных совпадений моделей в наборах из n_ϕ и n_ψ моделей, A - система пар подмножеств множества моделей Ω, образующая σ- алгебру событий, и p- вероятность на $(\Omega 1, A)$.

Определим на $\Omega1\times A$ случайную величину ξ так: $\xi(k)=\rho_k(B_i,B_j)$, $k=0,1,2,\ldots,\min(n_\phi,n_\psi)$, т.е. ξ каждому kиз $\Omega1$ ставит в соответствие расстояние $\rho_k(B_i,B_j)$.

Вероятность появления этого события- расстояния для k на классе из n моделей можно вычислить так:

$p_k=\dfrac{C_n^k C_{n-k}^{n_\phi-k} C_{n-n_\phi}^{n_\psi-k}}{C_n^{n_\phi} C_n^{n_\psi}}$, где $C_n^{n_\phi}$ - число способов выбрать n_ϕ моделей из n моделей.

Действительно, так как любое множество (набор моделей), состоящее из n_ϕ моделей , может сочетаться с любым набором моделей, состоящим из n_ψ моделей, то число $(C_n^{n_\phi}\cdot C_n^{n_\psi})$ - количество всех способов выбрать два набора моделей , один из которых состоит из n_ϕ моделей, а другой из n_ψ моделей.

Выбрать k моделей, которые будут общими в этих наборах, из n моделей можно C_n^k способами. Тогда остальные $(n_\phi-k)$ и $(n_\psi-k)$ моделей в наборах должны быть дизъюнктными. Следовательно, остальные $(n_\phi-k)$ моделей для пополнения

набора, состоящего из k до n_ϕ моделей, можно выбрать $C_{n-k}^{n_\phi-k}$ способами, а

$(n_\psi-k)$ моделей для получения набора, состоящего из n_ψ моделей, с учетом наших предположений, -- $C_{n-n_\phi}^{n_\psi-k}$ способами.

Значит, имеется всего $\left(C_n^k C_{n-k}^{n_\phi-k} C_{n-n_\phi}^{n_\psi-k}\right)$ способов выбрать два набора, один из которых состоит из n_ϕ моделей, а другой из n_ψ, и в точности k моделей в этих наборах моделей совпадают. Поэтому вероятность того, что k моделей совпадет в наборах из n_ϕ и n_ψ элементных моделей, будет

равна $p_k=\dfrac{C_n^k C_{n-k}^{n_\phi-k} C_{n-n_\phi}^{n_\psi-k}}{C_n^{n_\phi} C_n^{n_\psi}}$.

В результате получим, что расстояния $\rho_k(B_i,B_j)=\dfrac{n_\phi+n_\psi-2k}{n}$ будут

появляться с вероятностями $p_k=\dfrac{C_n^k C_{n-k}^{n_\phi-k} C_{n-n_\phi}^{n_\psi-k}}{C_n^{n_\phi} C_n^{n_\psi}}$, где $k=0,1,2,\ldots,\min(n_\phi,n_\psi)$.

Предлагается эти вероятности или ближайшие к ним числа использовать в качестве весов расстояний для получения результирующего расстояния для данных вероятностных формул. Более подробно об этом будет немного дальше. Заметим, что при таком подходе главную роль играют не сами формулы, а числа определяющие количество моделей и их пересечения. Не имея другой информации, мы посчитали все интересующие нас подмножества для вычисления частости (или вероятности) появления расстояния для конкретного k. Используя свойство инвариантности расстояний между формулами и вероятностей высказываний (формул) [Лбов, Старцева,1999 -- Викентьев, Лбов, 1998], можно только что проведенные рассуждения проделать с меньшим носителем знаний: включающем только те модели теории, которые включают носители встречающиеся в двух формулах, для которых ищется расстояние. Будем считать, что мы так сделали с самого начала. И тогда подсчет в этом случае будет оптимальным.

Зная вероятности p_k для каждого расстояния $\rho_k(B_i, B_j)$, в качестве расстояния между вероятностными высказываниями $B_i = \left\langle \phi, p_\phi \right\rangle$ и $B_j = \left\langle \psi, p_\psi \right\rangle$ можно взять, например, наиболее вероятное $\rho(B_i, B_j) = \rho_m(B_i, B_j)$, где $p_m = \max\limits_k p_k$. Для такого $\rho(B_i, B_j)$ справедлива теорема 1. Для получения другого расстояния можно взять усредненное некоторого подмножества (отобранных по предпочтению) из полученных.

Более общо, беря произвольные p_k (исходя из экспертных оценок или дополнительных сведений экспертов, которые могут и не совпадать) для расстояний $\rho_k(B_i, B_j)$ так, чтобы получался закон распределения, получим самый общий случай для адаптивного поиска нужного расстояния между формулами с вероятностями.

Тогда в качестве расстояния между вероятностными высказываниями $B_i = \left\langle \phi, p_\phi \right\rangle$ и $B_j = \left\langle \psi, p_\psi \right\rangle$ можно взять величину, равную математическому ожиданию (центру тяжести) или среднему значению случайной величины ξ, т.е.

$\rho(B_i, B_j) = M\xi = \sum\limits_{k=0}^{\min(n_\phi, n_\psi)} \rho_k(B_i, B_j) \cdot p_k$. Для так введенного расстояния справедлива следующая

Теорема 2. Для расстояния $\rho(B_i, B_j)$ справедливы свойства:

$$0 \le \rho(B_i, B_j) \le 1.$$

$$\rho(B_i, B_j) = \rho(B_j, B_i).$$

$$\rho(B_i, B_j) \le \rho(B_i, B_s) + \rho(B_s, B_j).$$

Если $\rho(B_i, B_j) = 0$, то $B_i \equiv B_j$.

$$\rho(B_i, B_j) = 1 - \rho(B_i, \neg B_j) = \rho(\neg B_i, \neg B_j).$$

$$\rho(B_i, B_j) = \rho(B_i \wedge B_j, B_i \vee B_j).$$

Доказательство. Докажем сначала свойство 3. Для расстояний на $B_i = \left\langle \phi, p_\phi \right\rangle$ по свойству 3 теоремы 1 имеем $\rho_k(B_i, B_j) \le \rho_k(B_i, B_s) + \rho_k(B_s, B_j)$. Тогда по свойствам математического ожидания {а) $\xi \le \eta \Rightarrow M\xi \le M\eta$, б) $M(\xi + \eta) = M\xi + M\eta$ и получаем требуемое свойство для расстояния.

Докажем свойство 4. Пусть $\xi(k) = \rho_k(B_i, B_j) \ge 0$ и $M\xi = 0$. Тогда по свойству математического ожидания $\xi(k) = \rho_k(B_i, B_j) = 0$ с вероятностью равной 1. Тогда по свойству 4 теоремы 1 $B_i \equiv B_j$. Докажем свойство 5. Так как для расстояния $\rho_k(B_i, B_j)$ по свойству 6 теоремы 1 справедливо равенство $\rho_k(B_i, B_j) = 1 - \rho_k(B_i, \neg B_j) = \rho_k(\neg B_i, \neg B_j)$, тогда по свойствам математического ожидания {а) если $p(\xi = \eta) = 1$ и $\exists\ M\xi$, то $M\xi = M\eta$, б) $M(a + b\xi) = a + bM\xi$} получаем требуемое свойство для расстояния $\rho(B_i, B_j)$. Остальное доказывается аналогично.

Заключение

В работе предложены способы введения метрик на высказываниях экспертов -- формулах ИВ с вероятностями с помощью моделей логической теории исследуемой области. Исследование найдет применение в решении задач согласования вероятностных высказываний экспертов, кластеризаци, и в построении баз знаний и экспертных систем. Результаты верны для формул над бесконечными носителем, переносятся на формулы с переменными языка 1-го порядка и формулы от разнотипных переменных с использованием измеримых подклассов (для фиксированной неполной теории) измеримых (в том числе и метрических) моделей.

Библиография

[Блощицын В.Я.., Лбов, 1987] Блощицын В.Я., Лбов Г.С. О мерах информативности логических высказываний.// Доклады Республиканской Школы-Семинара «Технология разработки экспертных систем». Кишинев, 1987, с.12-14.

[Лбов, Старцева,1999] Лбов Г.С., Старцева Н.Г. Логические решающие функции и вопросы статистической устойчивости решений. Новосибирск: Издательство Института математики, 1999, 212 с.

[Vikent'ev, Lbov, 1997] Vikent'ev A.A., Lbov G.S.. Setting the metric and informativeness on statements of experts. // Pattern Recognition And Image Analysis. 1997, v. 7 (2), p. 175-189.

[Викентьев, Лбов, 1998] Викентьев А.А., Лбов Г.С. О метризациях булевой алгебры предложений и информативности высказываний экспертов. // Доклады РАН, 1998, т. 361 (2), с.174-176

[Бериков, 2009] Бериков В.Б. Кластерный анализ с использованием коллектива деревьев решений // Научный вестник НГТУ. 2009. № 3 (36). С.67-76.

[Лбов, Бериков, 2005] Лбов Г.С., Бериков В.Б. Устойчивость решающих функций в задачах распознавания образов и анализа разнотипной информации. Новосибирск: Изд-во Ин-та математики, 2005. - 218 с.

[Ершов, Палютин, 2004] Ершов Ю.Л., Палютин Е.А. Математическая логика. М.: Наука,2004, 336 с.

Authors' Information

Александр А. Викентьев – с.н.с., канд.физ-мат.наук , Институт математики СО РАН, пр. Академика Коптюга, д.4, Лаборатория анализа данных. Доцент, Новосибирский госуд. университет; e-mail: vikent@math.nsc.ru

Руслан А. Викентьев – инженер, Институт математики СО РАН, пр. Академика Коптюга, д.4, Лаборатория анализа данных. Ассистент-преподаватель, Новосибирский госуд. университет; e-mail: (ruslan.vikentiev@gmail.com)