# ЗАДАНИЕ РАССТОЯНИЙ И МЕР ИНФОРМАТИВНОСТИ НА ПРЕДИКАТНЫХ ЗНАНИЯХ ЭКСПЕРТОВ ДЛЯ КЛАСТЕР-АНАЛИЗА

А.А. Викентьев, Л.Н. Коренева

Институт математики СО РАН, Новосибирский госуниверситет
пр. Коптюга, 4, Новосибирск, Россия, 630090
vikent@math.nsc.ru

## АННОТАЦИЯ

Рассматриваются логические высказывания экспертов как предикатные формулы об иерархических объектах и предлагаются способы задания на таких высказываниях метрики и меры информативности. Исследование найдет применение к вопросам построения решающих функций распознавания образов, разработки экспертных систем и кластер-анализа. Предложенные функции информативности удовлетворяют всем естественным предъявляемым к ней требованиям.

## ВВЕДЕНИЕ

К настоящему времени достаточно хорошо развиты теория и методы построения решающих функций распознавания образов на основе анализа эмпирической информации, представленной в виде таблиц данных. Параллельно этому проявляется все больший интерес к построению решающих функций на основе анализа экспертной информации, заданной в виде логических ‘’знаний’’ нескольких экспертов [1-4,13]. При этом возникает задача согласования высказываний экспертов об иерархических объектах, а также задачи введения расстояния на таких “знаниях” и определения информативности этих “знаний”.

При решении задач распознавания образов, кластерного и регрессионного анализа важную роль играет информация, полученная от экспертов. В трудноформализуемых областях исследований (где используется иерархическое описание объектов) особую важность приобретают методы обработки эмпирической информации, представленной набором неполных (частичных) экспертных “знаний” (“знания” могут быть частично или полностью противоречивы). Очевидно, высказывания (“знания”) могут различаться по количеству содержащейся в них информативности. Информативность отражает важность сообщенной экспертом информации.

Работа является естественным продолжением работ [2-4], и знакомство с ними предполагается. В работе рассматриваются частично заданные высказывания экспертов об иерархических объектах, записанные в виде логических предикатных формул. Предлагаются способы задания метрики на таких высказываниях экспертов и меры информативности этих предикатных формул. Для введения метрики на высказываниях экспертов используется теоретико-модельный подход [7-8,10]. Изучаются свойства введенных метрик и связанных с ними мер. Полученные результаты апробированы на конференциях [5-6, 13-15].

## РАССТОЯНИЯ

Зафиксируем язык первого порядка $L$, состоящий из конечного числа предикатных символов. Пусть $A_n$ - непустое множество (элементы которого суть математические

объекты, то есть измерения переменных) мощности $n$ ($\leq n$). Во многих приложениях имеет смысл рассматривать модели с конечным множеством (носителем).

**Определение 1 .**[7-8]. Под интерпретацией $\gamma$ будем понимать отображение, ставящее в соответствие всем сигнатурным символам $\Omega$ языка $L$, а именно $\Omega = \left\langle P_1^{n_1}, P_2^{n_2}, \ldots \right\rangle$ (где $n_i$- местность предиката $P_i$), набор конкретных предикатов $P_i^{A_n}$, определенных на множестве $A_n$. Это позволяет говорить о модели $\left\langle A_n; \Omega^{A_n} \right\rangle$ сигнатуры $\Omega$. Здесь мы будем рассматривать модели только конечной сигнатуры.

Пусть имеется конечное число экспертов. Модели (в смысле теории моделей) задаются самими экспертами. Каждый эксперт задает свою интерпретацию $\gamma\left(P_i^{n_i}\right)$ символам языка в соответствующие частичные отношения (предикаты) на множестве $A_n$. То есть имеем частично заданные "знания" экспертов (записанные в виде формул).

Обозначим через $Mod_n(L)$ множество всех моделей языка $L$, определенных на множестве $A_n$ экспертами.

Введем расстояние на множестве частично заданных моделей. Модели различаются интерпретациями.

Определим расстояние между формульными подмножествами (предикатами) в каждой модели $M_i \in Mod_n(L)$, как меру их симметрической разности.

**Определение 2 .** [5,14-15]. Расстоянием между предикатами $P_k^{M_i}$ и $P_j^{M_i}$, частично определенными в модели $M_i$, назовем величину

$$\rho_{M_i}\left(P_k^{M_i}, P_j^{M_i}\right) = \mu\left(\widetilde{P}_k^{M_i} \Delta \widetilde{P}_j^{M_i}\right).$$

**Замечание.** Если рассматриваемые предикаты имеют разную арность, т.е. эксперт не высказывается о значении некоторой характеристики, тогда полагаем, что эта характеристика является пустым множеством.

Расстояние между предикатами на всем семействе моделей $Mod_n(L)$ определим как среднее на множестве расстояний в моделях.

**Определение 3 .** [5,15-16]. Расстоянием между предикатами $P_k$ и $P_j$, частично определенными на множестве $Mod_n(L)$, назовем величину

$$\rho_1\left(P_k, P_j\right) = \frac{\sum_{M_i \in Mod_n(L)} \rho_{M_i}\left(P_k^{M_i}, P_j^{M_i}\right)}{\left|Mod_n(L)\right|}.$$

Далее, для простоты обозначений, знак $\widetilde{\ }$ над предикатами и формулами будем опускать, когда это не вызывает недоразумений.

Теперь рассмотрим способ определения расстояния между предложениями.

Обозначим через $Mod(\phi)$ множество моделей из $Mod_n(L)$, на которых истинно предложение $\phi$, т.е. $Mod(\phi) = \left\{M_i \in Mod_n(L) \mid M_i \models \phi\right\}$.

Очевидно, существуют такие модели, на которых (не тавтологичное) предложение истинно, и такие, на которых оно ложно. Естественно измерять различие информации, содержащейся в предложениях, количеством моделей, на которых предложения принимают разные значения истинности.

**Определение 4 .**[13]. Расстоянием между предложениями $\phi$ и $\psi$ назовем величину

$$\rho_2(\phi,\psi)=\frac{\left|Mod\left((\phi\wedge\neg\psi)\vee(\neg\phi\wedge\psi)\right)\right|}{\left|Mod_n(L)\right|}.$$

Рассмотрим еще один способ определения расстояния между формулами. Дополним язык $L$ первого порядка константами из множества $M = Mod_n(L)$. Для этого множества $M$ рассмотрим произвольные кортежи $\bar{a}$ длины местности формул, равной $l(\bar{a})$. При подстановке кортежей в формулы, в предположении, что формулы имеют одинаковую местность (как этого добиться, было показано выше), формулы становятся предложениями.

**Определение 5 .** [5,14]. Расстоянием между формулами $\phi$ и $\psi$ назовем величину

$$\rho_3(\phi,\psi)=\min_{\bar{a}\in M^{l(\bar{a})}}\rho_2(\phi(\bar{a}),\psi(\bar{a})).$$

Доказана следующая теорема, из которой следует, что предложенные расстояния действительно являются метриками. В теореме доказаны и некоторые дополнительные свойства введенных расстояний.

**Теорема 1 .** Для любых формул ("знаний" экспертов) $\phi$, $\psi$, $\chi$ и для любой функции $\rho_i$ справедливы следующие свойства:

1. $0 \le \rho_i(\phi, \psi) \le 1$.
2. $\rho_i(\phi, \psi) = \rho_i(\psi, \phi)$ (симметричность).
3. Если $\rho_i(\phi, \psi) = \rho_i(\phi_1, \psi_1)$ и $\rho_i(\phi_1, \psi_1) = \rho_i(\phi_2, \psi_2)$, то $\rho_i(\phi, \psi) = \rho_i(\phi_2, \psi_2)$ (транзитивность).
4. $\rho_i(\phi, \psi) \le \rho_i(\phi, \chi) + \rho_i(\chi, \psi)$ (неравенство треугольника).
5. $\phi \equiv \psi \Leftrightarrow \rho_i(\phi,\psi) = 0$ ($\phi \equiv \psi$ здесь и далее обозначает эквивалентность формул относительно всех моделей экспертов, то есть для любого эксперта $i$ (задающего модель $M_i$) верно $\phi^{M_i} = \psi^{M_i}$).
6. $\phi \equiv \neg\psi \Rightarrow \rho_i(\phi,\psi) = 1$.
7. $\rho_i(\phi, \psi) = 1 - \rho_i(\phi, \neg\psi) = \rho_i(\neg\phi, \neg\psi)$.
8. $\rho_i(\phi, \psi) = \rho_i(\phi \wedge \psi, \phi \vee \psi)$.
9. $\rho_i(\phi, \neg\phi) = \rho_i(\phi, \psi) + \rho_i(\psi, \neg\phi)$.

Доказательство теоремы следует из определений, свойств вероятностной меры, теоретико-модельных и логических вычислений.

## МЕРЫ ИНФОРМАТИВНОСТИ

С точки зрения важности информации, сообщенной экспертом, естественно считать, что информативность высказывания тем выше, чем меньше моделей (или мера), на которых оно выполнимо, поэтому введем информативность следующим образом.

**Определение 6 .** [5-6,14]. Пусть $P$ - предикат, отражающий знание эксперта, тогда мерой информативности предиката $P$ назовем величину $I_i(P) = \rho_i(P, 1)$, где 1 -- тождественно истинный предикат, то есть $\bar{x} = \bar{x}$.

Для введенных расстояний получаем :

$$I_i(P)=\begin{cases}\dfrac{\sum\limits_{M_i\in Mod_n(L)}\mu\left(\neg P^{M_i}\right)}{\left|Mod_n(L)\right|}, \text{если расстояние } \rho_1\\ \dfrac{\left|Mod\left(\neg P\right)\right|}{\left|Mod_n(L)\right|}, \text{если расстояние } \rho_2\\ \dfrac{\min\limits_{\bar a\in M^{l(\bar a)}}\left|Mod\left(\neg P(\bar a)\right)\right|}{\left|Mod_n(L)\right|}, \text{если расстояние } \rho_3\end{cases}$$

Доказана следующая теорема:

**Теорема 2** . Для любых формул (“знаний” экспертов) $\phi$, $\psi$ и любого $\rho_i$ справедливы следующие утверждения

1. $0\le I_i(\phi)\le 1$.
2. $I_i(1)=0$.
3. $I_i(0)=1$.
4. $I_i(\phi)=1-I_i(\neg\phi)$.
5. $I_i(\phi)\le I_i(\phi\wedge\psi)$.
6. $I_i(\phi)\ge I_i(\phi\vee\psi)$.
7. $I_i(\phi\wedge\psi)=\rho_i(\phi,\psi)+I_i(\phi\vee\psi)$.
8. Если $\phi\equiv\psi$, то $I_i(\phi)=I_i(\psi)$.
9. Если $\rho_i(\phi,\psi)=0$, то $I_i(\phi\wedge\psi)=I_i(\phi\vee\psi)=I_i(\phi)$.
10. $I_i(\phi\wedge\psi)=\dfrac{I_i(\phi)+I_i(\psi)+\rho_i(\phi,\psi)}{2}$.
11. $I_i(\phi\vee\psi)=\dfrac{I_i(\phi)+I_i(\psi)-\rho_i(\phi,\psi)}{2}$.

Для доказательства теоремы используются введенные определения, доказанные выше свойства метрики и теоретико-модельные вычисления.

**ЗАКЛЮЧЕНИЕ**

Аналогичные результаты справедливы для моделей с мощностями не превосходящими фиксированного натурального числа $n$. Полученные результаты можно использовать для нахождения усредненных расстояний и информативностей предикатных знаний экспертов. Все результаты остаются справедливыми при предположении, что используемая модель счетна, а число экспертов конечно. Планируются применения и перенесение этих результатов на произвольные бесконечные области с измеримым классом моделей и счетное число экспертов. Результаты будут использованы для кластер-анализа структурированных знаний экспертов.

Работа выполнена при финансовой поддержке РФФИ № 98-01-00673 и программы ‘’Интеграция’’ Новосибирского государственного университета.

**ЛИТЕРАТУРА**


1. Блощицын В.Я., Лбов Г.С. О мерах информативности логических высказываний.// Доклады Республиканской Школы-Семинара “Технология разработки экспертных систем”. Кишинев, 1978, с.12-14.
2. Лбов Г.С., Старцева Н.Г. Логические решающие функции и вопросы статистической устойчивости решений. Новосибирск: Издательство Института математики, 1999, 212 с.
3. 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.
4. Викентьев А.А., Лбов Г.С. О метризациях булевой алгебры предложений и информативности высказываний экспертов. // Доклады РАН, 1998, т. 361 (2), с.174-176.
5. Викентьев А.А., Коренева Л.Н. Три способа задания расстояний на высказываниях экспертов.// Сборник научных статей Международной конференции ’’Компьютерный анализ данных и моделирование’’. Минск, 1998. С.160-166.
6. Викентьев А.А., Коренева Л.Н. К вопросу о расстояниях между формулами, описывающими структурированные объекты. // Математические методы распознавания образов (ММРО-99). РАН ВЦ, Москва, 1999. С.151-154.
7. Chang C.C., Keisler H.J. Model Theory. Studies in Logic and Foundations of Mathematics. 1973,v.73, 550 p.
8. Ершов Ю.Л., Палютин Е.А. Математическая логика. М.: Наука, 1991, 336 с.
9. Лбов Г.С. Методы обработки разнотипных экспериментальных данных. Новосибирск: Наука , 1981, 158 с.
10. Гончаров С.С. Счетные булевы алгебры и разрешимость. Новосибирск. Научная книга, 1996, 362 с.
11. Gaifman H. Concerning measures in the first order calculi. // Israel Journal of Mathematics, v. 2 (1), 1964, p. 1-18.
12. Fagin R. Probabilities on finite models. // The Journal of Symbolic Logic, v. 41 (1), 1976, p. 50-58.
13. Викентьев А.А. Задание метрик на высказываниях экспертов и их информативности. // Математические методы распознавания образов, ММРО–7. Москва, 1995, с. 14-16.
14. Коренева Л.Н. Применение различных метрик для оценки информативности высказываний экспертов. // Материалы XXXIV Международной научной студенческой конференции. Новосибирск, 1996, с. 40-41.
15. Коренева Л.Н. Задание метрик на высказываниях экспертов. // Материалы XXXVI Международной научной студенческой конференции. Новосибирск, 1998, с. 63.