

## ПОНЯТИЕ НЕСТРОГОЙ ВЕРОЯТНОСТИ В ИНДУКТИВНОМ ВЫВОДЕ

*Л.В. Аршинский, В.С. Лебедев*

ФГБОУ ВО «Иркутский государственный университет путей сообщения»

Сегодня анализ больших данных – одно из ведущих направлений в обработке информации. Накопленные на электронных носителях сведения становятся источником новых знаний, и добыча их из информационных массивов представляет большой прикладной интерес.

Существуют различные подходы к анализу – статистические методы, OLAP-анализ, методы распознавания образов (задачи сегментации и классификации) и т.д. Естественное место в этом ряду занимает индуктивная логика, которая предназначалась для вывода общих закономерностей из частных случаев. В первую очередь – это закономерности вида «Если..., то...».

Одними из основных методов индуктивного вывода являются метод единственного сходства, метод единственного различия и объединённый метод сходства и различия. Фактически они сводятся к анализу таблиц совместной встречаемости:

	$a_1$	$a_2$	...	$a_n$	$b$
1	$a_{11}$	$a_{21}$	...	$a_{n1}$	$b_1$
2	$a_{12}$	$a_{22}$	...	$a_{n2}$	$b_2$
3	$a_{13}$	$a_{23}$	...	$a_{n3}$	$b_3$
...	...	...	...	...	...
$K$	$a_{1K}$	$a_{2K}$	...	$a_{nK}$	$b_K$

Здесь  $a_i$  и  $b$  – наблюдаемые явления, между которыми ищется связь типа  $a_i \rightarrow b$  (вид импликации может быть более сложным). Факты наличия/отсутствия явлений рассматривается двузначным (булевым) образом, когда  $a_{ik}, b_k \in \{0,1\}$  ( $k=1 \dots K$ ,  $K$  – число опытов). Каждая строка таблицы – это наблюдение, которое можно положить в основу заключения об импликации. Если требованию совместной встречаемости удовлетворяет только часть строк таблицы, можно использовать (и используются) статистико-вероятностные представления, когда подсчитывается доля «правильных» опытов в общем количестве. Это позволяет говорить о «вероятности» (достоверности) импликации  $a_i \rightarrow b$ . Однако при этом не учитывается возможность неполноты и противоречивости информации о явлениях. Например, когда утверждения об их наличии/отсутствии поступают из различных источников, которые могут противоречить друг другу и/или быть малоубедительными.

При наличии противоречащих и малоубедительных свидетельств более перспективным выглядит использование вероятностно-статистических представлений, использующих понятие нестрогой вероятности. Оно основано на векторной семантике, когда истинность утверждений (в данном случае – утверждений о наблюдении явления) описывается вектором ⟨Истина; Ложь⟩. Использование нестрогих вероятностей делает вывод устойчивым к низкому качеству (убедительности, согласованности) входных данных.

Анализ таблиц совместной встречаемости можно использовать для автоматизированного пополнения продукционных баз знаний. Источником продукции при этом служат большие данные.