

ФОРМАЛИЗАЦИЯ ПОНЯТИЯ ПРЕДСКАЗАНИЕ

Витяев Е.Е.¹, Харламов Е.Ю.²

¹Институт математики им. С.Л.Соболева, Новосибирск, Россия,
vityaev@math.nsc.ru

²Новосибирский государственный университет, Новосибирск, Россия,
is-3@gorodok.net

1 Введение

Основная тема работы - формализация и исследование понятия предсказание. Предсказание – важный элемент деятельности человека, связанной с познанием законов окружающего мира и дальнейшим использованием их для получения предсказаний. Первоначально исследование направленно на накопление фактов, данных. После этого факты систематизируются, обнаруживаются закономерные связи. Дальнейшее прогнозирование характера развития событий, основано на предсказании, построенном по закономерностям.

Авторами рассмотрены две формализации понятия предсказание, основанные на существующих работах по дедуктивным и индуктивным выводам и предложена новая формализация, основанная на семантическом вероятностном выводе.

2. Существующие подходы.

Первый подход (дедуктивный) предложен Карлом Поппером [1], [2]. Предсказание по Попперу – дедуктивный вывод, основанный на универсальном законе и сингулярных высказываниях. Результатом предсказания является сингулярное высказывание, дедуцированное по посылкам.

Общая схема дедуктивного вывода:

$$\begin{array}{l} U \\ I \\ \hline E \end{array} \left. \begin{array}{l} \} \text{Посылки или гипотезы (составляющие объясняющее)} \\ \\ \} \text{Заключение (являющееся объясняемым)} \end{array} \right\} (1)$$

U - общий закон или набор универсальных высказываний, то есть гипотез, носящих характер естественных законов.

I - набор сингулярных высказываний, которые относятся только к специфическому обсуждаемому событию; они есть суть измерения, наблюдения, описывающие контекст исследуемой ситуации. Начальные условия по Попперу.

E - сингулярное высказывание, называемое специфическим или сингулярным предсказанием.

Второй подход (индуктивный) предложен Карлом Гемпелем [3].

Индуктивно Статистическая (I-S) модель Гемпеля, основанная на вероятности, была построена для представления не дедуктивных выводов. Общая схема индуктивного вывода:

$$\begin{array}{l} P(G, F) = r \\ Fb \\ \hline Gb \end{array} [r] \quad (2)$$

F, G - свойства.

$P(G, F) = r$ - есть статистический закон, утверждающий, что относительная частота объектов s, имеющих свойство G (обозначается Gs) среди объектов, имеющих свойство F (обозначается Fs) равна r.

Fb – есть факт того, что объект b удовлетворяет свойству F (этот сингулярный факт- есть начальные условия по Попперу).

[r] указывает: какая степень индуктивной вероятности присуждается выводу (предсказанию) Gb, основанному на данных посылках.

3. Проблемы рассмотренных подходов.

Проблема подхода Поппера в том, что для предсказаний требуются достоверные не подвергающиеся сомнению (универсальные) законы, то есть законы, которые до сих пор не фальсифицированы (примером являются законы физики). Таким образом, из-за наложенного ограничения модель применима лишь для узкого класса задач, где от нас требуется дедуцировать: какое свойство будет (или объяснить наблюдаемое свойство) у конкретного объекта, являющегося частным случаем рассматриваемого класса, чье поведение подчиняется универсальному закону.

В тоже время, в рамках направлений Интеллектуальный Анализ Данных (Machine Learning) и Извлечение Знаний (Knowledge Discovery in Data Bases) теории не удовлетворяют требованиям Поппера, поскольку являются индуктивными.

При использовании подхода Гемпеля возникает, так называемая, проблема индуктивной двусмысленности. Суть проблемы в том, что для предсказания одного и того же сингулярного факта можно построить несколько индуктивных выводов по имеющимся данным с разной степенью индуктивной вероятности, присуждаемой заключению. Для выявления выводов пригодных для предсказания Гемпель ввел Правило Максимальной Определенности (или Специфицированности; кратко RMS), которое накладывает на множество объектов, выделяемых посылкой статистического закона, условие однородности относительно предсказываемого сингулярного факта. Ограничение однородности слишком сильно чтобы быть проверенным в условиях реально решаемых задач, по крайней мере, оно неприемлемо для широкого класса задач, где априорных знаний недостаточно для применения RMS. Поэтому для них в силу естественных обстоятельств мы не можем удовлетворить предъявленным I-S моделью требованиям. Кроме того, Гемпель ничего не говорит о том, как искать среди множества существующих индуктивных выводов тот, который удовлетворит RMS.

4. Предлагаемый подход

Авторами предлагается новый подход к формализации предсказания, определяемого как семантическое предсказание.

Общая идея следующая [4]: предсказание формализуется как индуктивный вывод по начальным условиям и не универсальному закону. Закон, пригодный для предсказания, называемый в работе удовлетворяющий Правилу Максимальной Вероятности (ПМВ), ищется при помощи семантического вероятностного вывода. Поэтому предложенная формализация радикально отличается от дедуктивного вывода тем, что законы в посылках рассматриваются не как универсальные, а как статистические. Отличие подхода от гемпелевского в принципиально ином правиле выбора индуктивного вывода, пригодного для предсказания.

Предлагается алгоритм, хоть и не детерминированный, нахождения требуемого вывода. Это, по меньшей мере, позволяет избежать нежелательного полного перебора данных, который вполне вероятен при гемпелевском подходе.

Для нахождения необходимых статистических законов используется представление сингулярных фактов через факты логической программы, статистических законов через правила логической программы, а сингулярных предсказаний через одноатомные запросы к логической программе. Сам процесс нахождения рассматривается, как

вероятностный вывод по вероятностной модели данных, не использующий правила логического вывода. Суть того, как выводится правило, удовлетворяющее ПМВ, следующая:

- поиск осуществляется путем движения вдоль “уточняющего” графа. В этом графе правила “уточняются” либо добавлением произвольного атома (или конъюнкции атомов) в посылку, либо применением подстановки. Выбор уточнения, удлиняющего соответствующую ветвь графа, определяется требованием увеличения условной вероятности, определяемой по вероятностной модели данных. Результатом вычисления является мажорантное наилучшее для предсказания правило, результирующая подстановка и достигнутая условная вероятность.

- На уточняющие правила в вероятностном выводе, согласно ПМВ, можно наложить дополнительное требование - чтобы каждый атом в посылке был ‘существенным’ для предсказания атома в заключении, то есть удаление любого из них уменьшает условную вероятность правила. Такие правила в работе называются вероятностными закономерностями (обозначим их множество через $PR(\mathcal{M})$).

Доказывается, что таким образом построенный индуктивный вывод предсказывает лучше любого другого (в смысле максимальности степени индуктивной вероятности) индуктивного вывода при одних и тех же данных.

Рассмотрим классический пример Гемпеля, на котором проиллюстрируем применение предлагаемого авторами подхода.

Нам нужно объяснение сингулярного факта: быстрое выздоровление Джона Джонса от инфекции стрептококка. Гемпель дает следующее объяснение:

$$P(G, F \wedge H) = r$$

$$\frac{Fb \wedge Hb}{Gb} [r] \quad (3)$$

Где F означает ‘болеть стрептококком’, H- ‘быть леченым пенициллином’, G- ‘быстрое выздоровление’, b означает самого Джона Джонса. r близко к 1.

Приведенное объяснение показывает (предсказывает) быстрое выздоровление Джона Джонса.

Предположим, что существуют устойчивые к лекарству бациллы стрептококка, и если некто инфицирован ими, то вероятность его быстрого выздоровления низка. Положим, что Джон Джонс болен именно такой инфекцией, тогда факт его быстрого выздоровления объясняется следующей схемой:

$$P(G, F \wedge H \wedge J) = r'$$

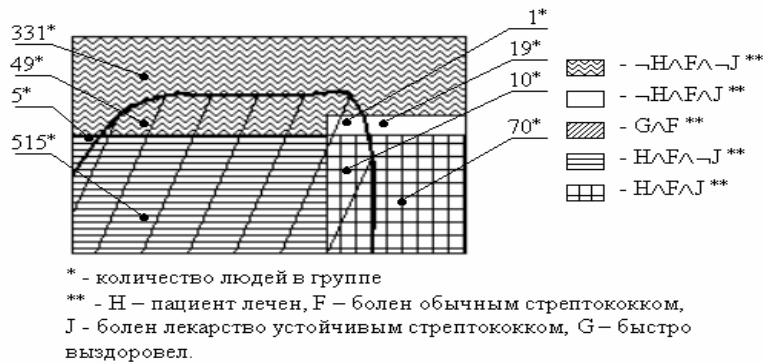
$$\frac{Fb \wedge Hb \wedge Jb}{Gb} [r'] \quad (4)$$

Где J означает ‘быть зараженным лекарством устойчивой инфекцией’, r' близко к 0.

Хорошо видна проблема, названная Гемпелем объяснительной (индуктивной) двусмысленностью: неясно какой из двух существующих выводов использовать для предсказания (объяснения) Gb.

При помощи изложенного авторами статьи подхода, вопрос о возможном быстром выздоровлении Джона Джонса решается следующим образом: пусть статистика людей болевших стрептококком (всего 1000 человек) имеет следующее графическое представление (рис 1):

Рис1:



ПРОЛОГ правила, которые можно составить, исходя из предложенных данных следующие:

$C_0: (G(s) = 1) \leftarrow (F(s) = 1), (H(s) = 1), (J(s) = 1); \mu(C_0) = 1/8.$

$C_1: (G(s) = 0) \leftarrow (F(s) = 1), (H(s) = 1), (J(s) = 1); \mu(C_1) = 7/8.$

$C_3: (G(s) = x) \leftarrow (F(s) = 1), (H(s) = 1), (J(s) = 1); \mu(C_3) = \min \{ \mu(C_0), \mu(C_1) \} = 1/8.$

...

Где s – переменная для больного стрептококком, x – бинарная переменная.

Множество фактов ($D(N)$): $(F(b) = 1) \leftarrow$, $(H(b) = 1) \leftarrow$, $(J(b) = 1) \leftarrow$, их меры - единицы.

Построив множество $PR(\mathcal{M})$, легко убедиться, что для запроса $\leftarrow (G(b) = 1)$ ('Выздоровеет ли быстро Джон Джонс?') мажорантное наилучшее для предсказания правило будет C' : $(G(s) = 1) \leftarrow (F(s) = 1), (H(s) = 1); \mu(C') = 7/8$, то есть ответ – да с вероятностью 7/8. Для запроса $\leftarrow (G(b) = 0)$ (('Верно ли, что Джон Джонс не выздоровеет быстро?')) имеем C'' : $(G(s) = 0) \leftarrow (F(s) = 1), (J(s) = 1); \mu(C'') = 0.89$, то есть ответ – нет с вероятностью 0.89. Для запроса $\leftarrow (G(b) = x)$ ('Что произойдет с Джоном Джонсом?') имеем C : $(G(s) = 0) \leftarrow (F(s) = 1), (J(s) = 1); \mu(C) = 0.89$, то есть ответ – быстрого выздоровления не будет с вероятностью 0.89. Приведенные результаты предсказаний можно получить, построив вероятностный вывод, используя правила из $PR(\mathcal{M})$.

5. Заключение

В заключении конкретизируем использованные в работе понятия.

ОПРЕДЕЛЕНИЕ 1. Вывод вида:

$$P(G, F_1 \wedge \dots \wedge F_n) = r$$

$$\frac{F_1 b \wedge \dots \wedge F_n b}{Gb} [r] \quad (5)$$

где $\{F_1'b, \dots, F_m'b\} \supseteq \{F_1b, \dots, F_nb\}$, $F_i'b$ ($i = 1, \dots, m$) – начальные данные, назовем удовлетворяющим *Правилу Максимальной Вероятности* (ПМВ) если

1. Любой вывод вида (5), предсказывающий Gb , имеет степень индуктивной вероятности $r' \leq r$.
2. Если удалить любое свойство F_i , $i = 1, \dots, n$ из посылок статистического закона, то получим статистический закон со степенью индуктивной вероятности меньшей r .

Сформулируем определение семантического предсказания.

ОПРЕДЕЛЕНИЕ 2. *Семантическое Предсказание* – индуктивный вывод вида (5) основанный на

а) не универсальном законе - мажорантном наилучшем для предсказания правилом полученным в результате вероятностного вывода по данным $D(N)$ из некой модели N и множеству вероятностных закономерностей с непустой посылкой;

б) начальных условиях (сингулярных высказываниях), являющихся элементами модели N , одной из множества данных G .

Результат семантического предсказания – сингулярное высказывание, индуцированное по посылкам.

Оценка семантического предсказания (обозначим η_s) – оценка условной вероятности мажорантного наилучшего для предсказания правила.

Основной результат:

ТЕОРЕМА 1. Если индуктивное предсказание Q (вида (2)) имеет оценку r , то семантическое предсказание, полученное при тех же данных, имеет оценку $\eta_s \geq r$.

Таким образом, предлагаемый подход имеет следующие преимущества перед существовавшими ранее:

1. Семантическое предсказание предсказывает лучше любого другого (в смысле максимальности степени индуктивной вероятности) индуктивного вывода при одних и тех же данных.
2. Семантическое предсказание решает проблемы, возникшие в ранее предложенных подходах.

Литература

1. К.Р.Поппер, Логика и рост научного знания // Москва, 'Прогресс', 1983.
2. К.Р.Поппер, Объективное знание, эволюционный подход// Москва, УРСС, 2002.
3. C.G.Hempel, Deductive-Nomological vs. Statistical Explanation. // Minnesota Studies in the Philosophy of Science III, University of Minnesota Press, Minneapolis, 1962.
4. Витяев Е.Е. Семантический подход к созданию баз знаний. Семантический вероятностный вывод наилучших для предсказания ПРОЛОГ-программ по вероятностной модели данных. // Логика и семантическое программирование (Вычислительные системы, вып. 146), Новосибирск, 1992, с.19-49.