МАТЕМАТИЧЕСКИЕ МОДЕЛИ
И ВЫЧИСЛИТЕЛЬНЫЕ СТРУКТУРЫ
(Вычислительные системы)
2004 год Выпуск 173
УДК 53.02.+519.7+519.812.2

КОНЦЕПЦИЯ ПРЕДСКАЗАНИЯ: ФОРМАЛИЗАЦИЯ И АНАЛИЗ[1]

Е.Ю. Харламов

Введение

Тема работы — формализация и исследование понятия предсказание (см. [10]).

Предсказание — важный элемент деятельности, связанный или направленный на познание законов окружающего мира либо на исследование некоторой предметной области в контексте влияния различных факторов и дальнейшее использование извлеченных знаний для достижения цели. Первоначально исследование направленно на наблюдение и накопление первичных фактов, данных, которые будут использованы для понимания явления. Затем факты систематизируются, обнаруживаются закономерные связи между различными фактами. Это делается, исходя из априорного предположения, что именно такие связи будут необходимы для достижения цели. Дальнейшее прогнозирование характера развития событий, основанное на предсказании, позволит выбрать оптимальный, наиболее эффективный алгоритм действий для достижения цели.

Заметим, что способность правильно предсказывать — существенная характеристика интеллектуальных систем, таких как экспертные системы, системы распознавания.

Процедуру предсказания можно условно представить [7] в виде схемы, приведенной на рис. 1.

[1] Работа поддержана грантом НШ-2112.2003.1.

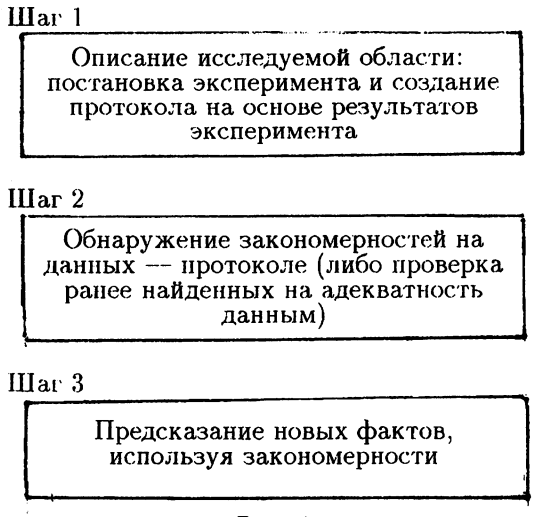


Рис. 1

Конкретные формализации отличаются друг от друга различным трактованием такого понятия как закономерность, способом извлечения закономерностей из данных, а также способом предсказания новых фактов по закономерностям. Под закономерностью можно понимать как универсальный закон природы, так и вероятностный, статистический закон. В свою очередь индивидуальный акт предсказания можно рассматривать либо как дедуктивный, либо как индуктивный вывод. Следующее различие заключается в способе извлечения закономерностей из данных.

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

Дедуктивный и индуктивный выводы являются составляющими так называемого внутритеоретического научного вывода, т.е. вывода, представленного внутри научных теорий и предназначенного для движения от одних данных (называемых причиной или посылками) к другим (называемым следствием или заключением). Причем такое движение осуществляется при по-

мощи теории или универсального (для дедуктивного вывода) либо статистического (для индуктивного вывода) закона, который является связующим звеном между причиной и следствием. Заметим, что такая классификация выводов основана на философии индукции и вероятности Рудольфа Карнапа [4].

Индуктивно–вероятностное представление внутритеоретических научных выводов в наиболее определенной форме было предложено Карлом Гемпелем через его модель научного объяснения. Дедуктивные выводы же представлены в работах Карла Р. Поппера [1,2].

Предложенная Гемпелем основанная на вероятности индуктивно–статистическая (I-S) модель для представления не дедуктивных выводов [5], однако, неспособна удовлетворительно решить, так называемую, проблему индуктивной двусмысленности, даже когда Гемпель вводит требование максимальной определенности, так называемое "Requirement of Maximal Specificity" (RMS). В результате обнаруживается проблема в представлении не дедуктивных выводов через численно–вероятностный подход, и, более того, это свидетельствует о несостоятельности индуктивно–вероятностного подхода к внутритеоретическим научным выводам в существующей на данный момент форме [6].

В работе предлагается иной подход к формализации предсказания, определяемого как семантическое предсказание.

Предсказание формализуется как индуктивный вывод по начальным данным и не дедуктивному закону. Закон, пригодный для предсказания, называемый в работе: удовлетворяющий требованию максимальной вероятности ("Requirement of Maximal Probability", RMP), ищется при помощи семантического вероятностного вывода. Поэтому предложенная формализация радикально отличается от дедуктивного вывода тем что законы в посылках рассматриваются не как тождественно истинные, но как статистические. Отличие подхода от не дедуктивных выводов Гемпеля в принципиально ином правиле выбора выводов пригодных для предсказания и тем самым в рамках предлагаемой формализации решается проблема индуктивной двусмысленности, возникающая с RMS. Для нахождения необходимых статистических законов воспользуемся представлением сингулярных

фактов, являющихся начальными условиями, через факты в ПРОЛОГ-программах, статистических законов через правила ПРОЛОГ-программы, а сингулярных предсказаний через одноатомные запросы к построенным ПРОЛОГ-программам. Сам процесс нахождения рассматривается, как вероятностный вывод по вероятностной модели данных, не использующий правила логического вывода. Доказывается, что таким образом построенный индуктивный вывод предсказывает лучше индуктивного вывода Гемпеля на одних и тех же фактах.

1. Дедуктивный подход

Рассмотрим формализацию, предложенную Карлом Р. Поппером [1]. Поппер считает, что дать причинное объяснение некоторого события или, что тоже, предсказать некоторое событие — значит, построить его дедуктивный вывод. Общая схема вывода следующая [2]:

$$\left.\begin{array}{l} U \\ \underline{I\quad} \end{array}\right\} \text{посылки или гипотезы (составляющие объясняющее)},$$
$$E \quad \} \text{заключение (являющееся объясняемым)}, \qquad (1)$$

где U — общий закон или набор *универсальных высказываний*, т.е. 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 (Поппер назвал такие сингулярные высказывания — *начальные условия*); E- сингулярное высказывание, называемое специфическим или сингулярным *предсказанием*.

Таким образом, *предсказание по Попперу* — дедуктивный вывод, основанный на универсальном законе и сингулярных высказываниях (начальных условиях). *Результатом предсказания* является сингулярное высказывание, дедуцированное из посылок.

Согласно Попперу, все предсказания можно разделить на три класса: высказывание о прошлом (ретросказание), объяснение имеющихся в настоящее время высказываний и предсказание как таковое, т.е. объяснение чего-то, что еще не произошло.

Заметим, что предсказывать мы можем только сингулярные факты, которые, в свою очередь, могут быть далее использованы, как начальные условия в посылках новых предсказаний. Откуда берутся универсальные высказывания, и что они из себя представляют? Поппер рассматривает их как элементы некой общей теории либо, как теорию как таковую, причем такого рода теории не выводятся из опыта прошлых наблюдений, а являются лишь нашей догадкой, предположением.

Заметим, что особенность любой теории, по Попперу, в том, что она может быть проверена независимо от данной рассматриваемой ситуации, т.е. от данных начальных условий в конкретном предсказании. Начальные условия в свою очередь проверяются независимо от теории, используемой в предсказании.

Рассмотрим пример: пусть U — универсальное высказывание постулирующее, что на все яблоки, находящиеся в атмосфере земли действует её сила притяжения, I — начальные условия, утверждающие про конкретное яблоко, что оно отсоединено от ветки дерева, на котором яблоко выросло. Из этих посылок мы можем дедуцировать факт падения яблока.

2. Индуктивный подход

Предсказание имеет вид *индуктивного вывода*

$$\left.\begin{array}{l} \mathrm{P(G,F)=r} \\ \underline{\mathrm{Fb}\qquad\qquad} \end{array}\right\} \text{посылки индуктивного вывода,}$$

$$\mathrm{Gb}\qquad [\mathrm{r}]\ \} \text{заключение индуктивного вывода,} \qquad (2)$$

где F, G — свойства.

$P(G,F) = r$ — есть статистический закон (в наших терминах закономерность, т.е. Гемпель понимает закономерность как статистический закон), утверждающий, что относительная частота объектов s, имеющих свойство G (обозначается Gs) среди объектов, имеющих свойство F (обозначается Fs) равна r. Иными словами,

$$P(G,F) = \frac{\sharp(x|G(x) \wedge F(x))}{\sharp(x|F(x))}.$$

Статистический закон есть суть импликация вида: $(Fx = 1) \longrightarrow (Gx = 1)$ с индуктивной вероятностью r.

Назовем r — индуктивной вероятностью статистического закона.

Назовем G — свойством из заключения статистического закона, а F — свойством из посылок статистического закона. Заметим, что F может представлять собой конъюнкцию свойств.

Fb — факт, что объект b удовлетворяет свойству F (начальные данные, элемент протокола измерений).

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

Каким образом находить закономерности? Гемпель не предлагает никаких алгоритмов поиска закономерностей, но утверждает, что их можно извлечь из данных.

В связи с этим возникает так называемая проблема индуктивной двусмысленности. Суть которой в том, что мы можем извлечь несколько статистических законов для предсказания одного и того же факта, т.е. законов с одним и тем же свойством в заключении. Извлеченные законы будут различаться свойствами из посылки закона и, возможно, степенью индуктивной вероятности (условной вероятностью). Таким образом, по одним данным возможно построение нескольких предсказаний одного и того же факта с различной степенью индуктивной вероятности.

Гемпель предлагает способ разрешения возникшей проблемы: автор считает, что предсказывать можно лишь при помощи максимально специфицированных (определенных) индуктивных выводов. Такие выводы он называет удовлетворяющими RMS. Суть RMS в следующем.

ОПРЕДЕЛЕНИЕ 1. Индуктивный закон в посылке индуктивного вывода максимально специфицирован, если при добавлении любых свойств в посылку закона индуктивная вероятность закона не меняется.

ОПРЕДЕЛЕНИЕ 2. Индуктивный вывод удовлетворяет RMS, если индуктивный закон в посылке максимально специфицирован.

Рассмотрим ниже способ решения проблемы индуктивной двусмысленности в терминах однородности.

Сначала сформулируем требование максимальной определенности в этих терминах.

По определению, унарный предикат на некотором множестве X есть некое подмножество $Y \subseteq X$ данного множества. Пусть G — предикатный символ (унарный), интерпретация которого в X есть Y. Пусть G_1 — предикатный символ (унарный), интерпретация которого в X есть Y_1. Заметим, что формулы $G(x)$ и $G_1(y)$ истины на множествах Y и Y_1 соответственно. Формула $G(x) \wedge G_1(y)$ истина на множестве $Y_3 = Y \cap Y_1$.

Заметим также, что статистический закон, т.е. закономерность, указывающая связь между свойствами G и G_1 для фиксированного элемента из X, будет иметь вид $P(G, G_1) = |Y_3|/|Y_1|$.

Переформулируем в терминах объект–свойство. Выборка составляет $|X|$ объектов, X — все объекты (элементы) выборки. После проведения процедуры измерения получен протокол, в котором отражены следующие данные: $|Y|$ объектов обладают свойством G и $|Y_1|$ — свойством G_1.

Будем говорить, что

ОПРЕДЕЛЕНИЕ 3. Множество объектов Y выделяется из множества объектов X свойством G при данном протоколе наблюдений (или, свойство G выделяет множество объектов Y из множества объектов X), если $Y \subseteq X$ и в X только у объектов из Y в результате эксперимента зафиксировано свойство G.

Ниже будет опускаться уточнение: при данном протоколе наблюдений, если это не приведет к неясности.

Далее имеем, что свойствами G и G_1 одновременно обладают $|Y_3|$ объектов. Если $Y_1 \subseteq Y$, то можно говорить, что свойство G выделяет объекты Y (те которые удовлетворяют G, их $|Y|$ штук) из выборки X, а свойство G_1 выделяет объекты Y_1 из уже выделенных объектов Y, при данном протоколе.

Пусть в протоколе отражены свойства $G, G_1, \ldots, G_n$.

Конъюнкция $(G_{i_1} \wedge \cdots \wedge G_{i_k})$ выделяет множество объектов Z из X, числом $|Z|$, $G_{i_j} \in \{G_1, \ldots, G_n\}$. Обозначим $(G_{i_1} \wedge \ldots \wedge G_{i_k}) = F$.

Конъюнкция $(G \wedge G_{i_1} \wedge \ldots \wedge G_{i_k})$ выделяет множество объектов Z_1 из X, числом $|Z_1|$.

Ниже отождествим свойство (предикатный символ) с множеством элементов для которых оно выполнено (с его конкретной интерпретацией на данном множестве), тогда имеем: $(G_{i_1} \wedge \ldots \ldots \wedge G_{i_k})$ отождествлено с Z, и $(G \wedge G_{i_1} \wedge \ldots \wedge G_{i_k})$ отождествлено с Z_1.

Заметим, что статистический закон будет иметь вид

$$P(G, G_{i1} \wedge \cdots \wedge G_{ik}) = P(G, F) = \frac{|Z_1|}{|Z|}.$$

Будем говорить, что

ОПРЕДЕЛЕНИЕ 4. Множество $\{x|G_{i_1}(x) \wedge \cdots \wedge G_{i_k}(x)\}$ однородно относительно множества $\{x|G(x)\}$ (или, что тождественно $G_{i_1} \wedge \cdots \wedge G_{i_k}$ однородно относительно G) если $\forall G_j \in \in \{G_1, \ldots, G_n\} \setminus \{G_{i_1} \ldots G_{i_k}\}$ верно

$$\frac{\sharp(x|G(x) \wedge F(x))}{\sharp(x|F(x))} = \frac{\sharp(x|G(x) \wedge F(x) \wedge G_j(x))}{\sharp(x|F(x) \wedge G_j(x))}.$$

То есть, в множестве F относительная частота встречания элементов из G такая же как и в любом его подмножестве $G_j^{'}$ выделяемом из F неким свойством G_j, таким что G_j не из конъюнкции F и $G_j \neq G$.

Проиллюстрируем это определение на примере приведенном на рис. 2, где $1 - F \wedge G$, $2 - F \setminus G$, $3 - F \wedge G_j \wedge G$, $4 - F \wedge G_j \setminus G$.

Таким образом, свойство однородности выражается так: $\frac{S_1}{S_1 + S_2} = \frac{S_3}{S_3 + S_4}$ где S_i — площадь i–го участка на рис. 2

Таким образом, для любого множества $G_j^{'}$ (подмножества F, получаемого при пересечении F и G_j) относительная частота встречаемости элементов из G равна относительной частоте встречаемости элементов из G во всем F.

Заметим, что свойство однородности множества объектов (свойства) относительно некого подмножества (другого свойства) есть понятие относительное, то есть оно зависит от того: какие свойства объектов отражены в протоколе. Таким образом

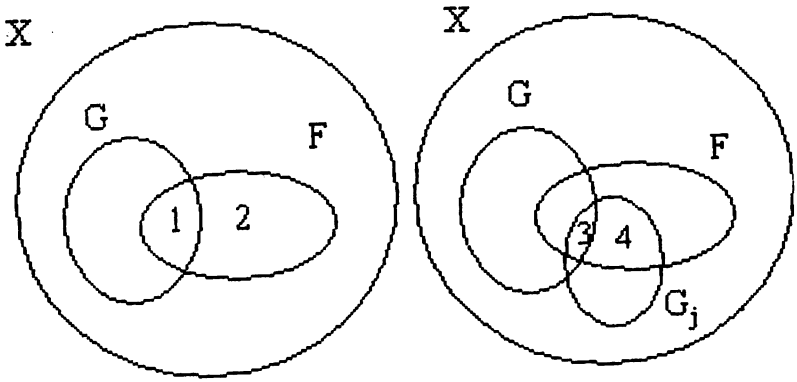


Рис. 2

для одного и того же набора объектов, но на разных протоколах (построенных для разных наборов свойств, или с иным числом пробелов в протоколе) один и тот же статистический закон может быть в одном случае однороден, а в другом — нет.

Подведем итог. В индуктивном подходе

1. Закономерность определяется, как статистический закон. Выделяется особая закономерность: максимально специфицированная закономерность (удовлетворяющая RMS).

2. Не предложены способы извлечения закономерностей из данных.

3. Способ предсказания новых фактов по закономерностям следующий.
   Выбирается статистический закон являющийся максимально специфицированным.
   Предсказание — индуктивный вывод, основанный на максимально специфицированном законе и фактах, свойства которых являются посылками статистического закона.

Рассмотрим классический пример Гемпеля. Нам нужно объяснение сингулярного факта: быстрое выздоровление Джона Джонса от инфекции стрептококка. У нас есть следующее объяснение:

$$\begin{array}{l} P(G, F \wedge H) = r, \\ \dfrac{Fb \wedge Hb}{Gb}[r], \end{array} \tag{3}$$

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

$$\begin{array}{l} P(G, F \wedge H \wedge J) = r', \\ \dfrac{Fb \wedge Hb \wedge Jb}{Gb}[r'], \end{array} \tag{4}$$

где J означает 'быть зараженным лекарство устойчивой инфекцией'. r' близко к 0.

Этот пример иллюстрирует проблему объяснительной или индуктивной двусмысленности. В рассмотренном примере с лекарство устойчивой инфекцией имеем два вывода Gb ((3) и (4)), причем посылки доказательств логически совместны, вывод один и тот же, но в первом случае решение сильно подкреплено посылками (следует из посылок с высокой вероятностью), а во втором сильно подрывается посылками (следует из посылок с вероятностью близкой к нулю). Суть проблемы в том, что неясно, какой из двух существующих выводов использовать для предсказания (объяснения) Gb.

3. Семантический подход

В этом параграфе предсказание определяется как *семантическое.*

ОПРЕДЕЛЕНИЕ 5. Семантическое предсказание есть вывод, удовлетворяющий RMP, вида:

$$\begin{array}{l} P(G, F_1 \wedge \cdots \wedge F_n) = r, \\ \dfrac{Fb_1 \wedge \cdots \wedge Fb_n}{Gb}[r], \end{array} \qquad (5)$$

где $\{F_1'b, \ldots, F_m'b\} \supseteq \{F_1b, \ldots, F_nb\}$, $F_i'b$, $i = 1, \ldots, m$, — данные, отраженные в протоколе; F_ib — начальные данные, остальные обозначения аналогичны индуктивному выводу Гемпеля.

В чем суть RMP?

ОПРЕДЕЛЕНИЕ 6 (см. [3]). Статистический закон с индуктивной вероятностью r, построенный по некоторому фиксированному протоколу, удовлетворяет RMP, если

1) любой иной статистический закон, построенный по данному протоколу и имеющий в заключении Gb, имеет индуктивную вероятность $r' \leq r$;
2) удалить любое свойство из посылок статистического закона, то получим статистический закон с индуктивной вероятностью меньшей r.

ОПРЕДЕЛЕНИЕ 7. Вывод вида (5) назовем удовлетворяющим RMP, если закон в посылке вывода удовлетворяет RMP.

Как найти вывод, удовлетворяющий RMP?

Для нахождения необходимого статистического закона используется представление сингулярных фактов через факты логической программы, статистических законов через правила логической программы, а сингулярных предсказаний через одноатомные запросы к логической программе. Переход в индуктивном выводе от посылок к заключению рассматривается как исполнение одноатомных запросов к построенным по индуктивным выводам логическим программам, т.е. как успешный SLDF-вывод.

Далее искать нужный статистический закон будем в терминах логического программирования. Будет найдено мажорантное наилучшее для предсказания правило, являющееся логическим аналогом закона удовлетворяющего RMP.

Общий вид логической программы с одним логическим правилом:

$$Pr\begin{cases} C_0 = G(x) \longleftarrow F_1(x), \ldots, F_n(x),\ n \geq 1, \text{ — правило логической программы,} \\ C_1 = F_1 b \longleftarrow \text{ — факт логической программы,} \\ \ldots \\ C_n = F_n b \longleftarrow \text{ — факт логической программы,} \\ \text{запрос} \longleftarrow G(b) \text{ — запрос к логической программе,} \end{cases}$$

где x — набор переменных, от которых зависит логическое правило C_0.

Искать будем при помощи *семантического вероятностного* [3] вывода (сравни с [8,9]).

ОПРЕДЕЛЕНИЕ 8. Семантический вероятностный вывод — пошаговый недетерминированный процесс построения логического правила, которое есть логический аналог статистического закона, удовлетворяющего RMP, т.е. построения мажорантного наилучшего для предсказания правила.

ОПРЕДЕЛЕНИЕ 9. Дерево семантического вероятностного вывода — ориентированное дерево с размеченными вершинами. Ориентация от корня к листьям. Изначально размечен лишь корень дерева, его пометка: предсказываемый атом либо без интерпретированных переменных, либо с интерпретированными переменными (т.е. предикатный символ атома — предсказываемое свойство). Далее дерево строится и параллельно размечается логическими правилами. Дерево считается построенным, когда закончена разметка его вершин, т.е. когда невозможно продолжать его построение ввиду невозможности продолжения процесса разметки.

Заметим, что интерпретированность переменных в помечающем корень атоме зависит от запроса: параметрический ли он или нет. Если запрос "верно ли, что свойство G у некоторого конкретного объекта b равно β (или лежит в неких рамках

для случая, когда значение b меняется непрерывно)", то запрос не параметризован и атом, помечающий корень без переменных. Если запрос "каково значение свойства G у некоторого конкретного объекта b (или в каких оно лежит рамках для случая, когда значение b меняется непрерывно)", то запрос параметризован и атом, помечающий корень с переменными.

На все правила, размечающие вершины дерева накладываются следующие условия.

1. Для любого правила, помечающего вершину дерева семантического вывода, должна существовать подстановка θ : $X \longrightarrow T$ (X — множество всех переменных, T — множество всех основных термов, в нашем случае, подстановка придает свойствам конкретные значения свойств из протокола для конкретного элемента из данного набора объектов), после применения которой к этому правилу полученное правило становится таковым, что для выполнения запроса (построенного как логический аналог предсказываемого свойства) достаточно применить последовательность только резолюций между фактами и этим правилом.

2. Каждый атом в посылке правила должен быть существенным для заключения правила, то есть удаление любого атома из посылок правила уменьшает вероятность правила.

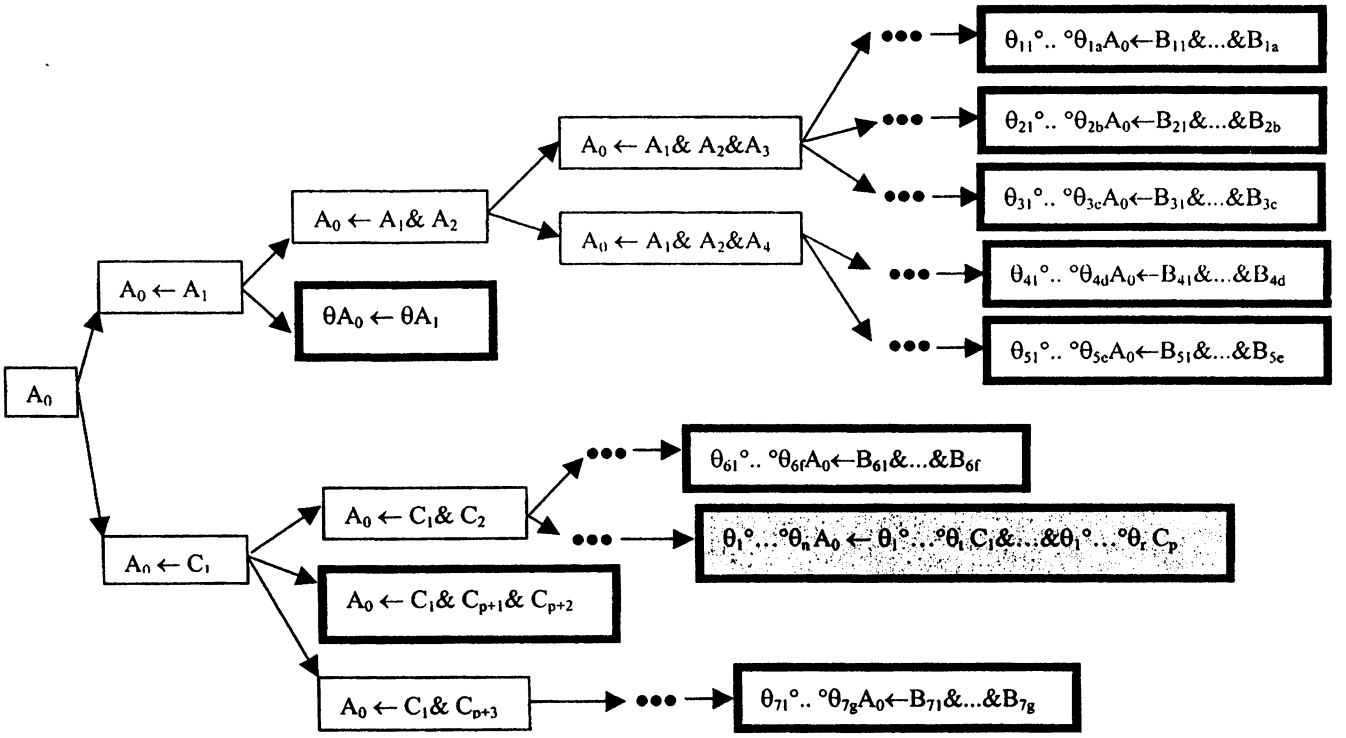
ОПРЕДЕЛЕНИЕ 10 (см. [3]). Назовем все правила удовлетворяющие условиям 1 и 2 вероятностными закономерностями.

Способ разметки.

Размечаем, двигаясь от корня к листьям последовательно по каждой из веток.

- На первом уровне от корня: вершины помечаются правилами с одним атомом в посылке и атомом, помечающим корень, в заключение правила.

- На i-ом ($i > 1$) уровне от корня: вершины помечаются правилами, полученными следующим образом: любая вершина метится правилом, отличающимся от правила, помечающего смежную вершину, лежащую ближе к корню,



где A_i, B_{ij}, C_i – атомы (атом: $A = P(t_1 \ldots t_n)$, где t_i – термы, P – предикатный символ);
θ, θ_i, θ_{ij} – подстановки ($\theta, \theta_i, \theta_{ij} : X \rightarrow T$, где X – множество переменных; T – множество термов);
■ - лист, в каждом листе – наилучшее для предсказания правило;
□ - промежуточная вершина,
■ - лист, в котором искомое правило: мажорантное наилучшее для предсказания правило.

Рис. 3. Уточняющее дерево

тем, что оно является его уточнением (определение понятия уточнение введем немного ниже).

- Процесс разметки останавливается, когда невозможно пометить ни одну вершину.

ОПРЕДЕЛЕНИЕ 11 (см. [3]). Логическое правило C_1 уточняет логическое правило C_2 если правило C_1 получено из правила C_2 одним из следующих способов:

1) добавлением произвольного атома (или конъюнкции атомов) в посылку;

2) применением подстановки,

а также верно, что вероятность C_1 строго больше вероятности C_2.

На рис. 3 приведен пример уточняющего дерева.

Для фиксированного протокола эксперимента для данного набора объектов сформулируем основные понятия описываемого метода предсказаний.

ОПРЕДЕЛЕНИЕ 12. Наилучшее для предсказания правило — это правило, находящееся в листе дерева семантического вероятностного вывода.

ОПРЕДЕЛЕНИЕ 13. Мажорантное наилучшее для предсказания правило — это наилучшее для предсказания правило, имеющее максимальную статистическую вероятность среди всех полученных семантическим вероятностным выводом.

Заметим, что при фиксированном протоколе эксперимента для данного набора объектов возможно существование нескольких мажорантных наилучших для предсказания правил.

Сформулируем основной результат работы. Приведена лишь схема доказательства, так как при его детальном изложении возникает необходимость формулировки и доказательства ряда дополнительных фактов, что невозможно из-за ограниченности размера статьи.

ТЕОРЕМА 1. *Если индуктивное предсказание Q (вида (5)) имеет оценку w, то семантическое предсказание, полученное при тех же начальных условиях, имеет оценку $\eta \geq w$.*

ДОКАЗАТЕЛЬСТВО. Пусть $P(G, F_1 \wedge \cdots \wedge F_n) = w$ — закон из посылок индуктивного предсказания Q, а $F_1 b, \cdots, F_n b$ —

начальные данные индуктивного предсказания Q. Тогда статистическому закону соответствует логическое правило $G(x) \longleftarrow F_1(x), \cdots, F_n(x)$ логической программы Pr, вероятность правила w, а начальным данным соответствует набор фактов $\longleftarrow F_1(b), \cdots, \longleftarrow F_n(b)$ логической программы Pr.

Поскольку Q есть индуктивное предсказание, то для построенной логической программы Pr существует успешный SLDF-вывод одноатомного запроса $G(x) \longleftarrow$. Отсюда мы заключаем существование набора подстановок $\theta_1, \cdots, \theta_n$, где $\theta_i : X \longrightarrow T$, X — множество переменных, T — множество основных термов. Причем $\theta_1 F_1(x), \cdots, \theta_1 \circ \cdots \circ \theta_n F_n(x)$ лежат в множестве фактов.

Построим размеченное дерево следующим образом. На первом шаге дерево состоит лишь из корня. Пометка корня: $G(x)$; $i := 1$.

На втором шаге добавим к корню n листьев, пометим их логическими правилами $G(x) \longleftarrow F_1(x), \cdots, G(x) \longleftarrow F_n(x)$. Каждую вершину — одним правилом и разные вершины - разными правилами; $i := 2$.

На шаге i: добавляем к листьям построенного на предыдущем шаге дерева по одному сыну и помечаем его. Пометка сына есть логическое правило с посылкой отличной от посылки логического правила помечающего его отца тем, что в нее добавляется один атом. Атом —любой из множества $\{F_1(x), \cdots, F_n(x)\} \setminus \{F_{j_1}(x), \cdots, F_{j_i}(x)\}$, где $F_{j_k}(x)$, $k \in \{1, \cdots, i\}$ все атомы из посылки логического правила помечающего отца; $i := i + 1$.

Процесс разметки останавливается когда $i = n$. На выходе имеем размеченное дерево D.

Далее выберем в дереве множество веток таких, что статистическая вероятность правила в третьей вершине дерева (корень считаем за первую вершину) больше чем вероятность правила во второй вершине. Получим множество R веток; $i := 1$.

Теперь в R выберем те ветки где статистическая вероятность правила в четвертой вершине больше чем в третьей; $i := 2$.

Выполнять описанный алгоритм модификации множества R до тех пор пока $i < n - 2$.

По окончанию работы алгоритма возможны две ситуации: R — пусто и R — не пусто.

Заметим, что в D ровно n веток длины $n+1$.

Если R пусто, то в процессе построения D, двигаясь по каждой из n возможных веток дерева, вероятность правил сначала росла (как минимум она возросла при переходе от первой ко второй вершине в каждой ветке), а затем, начиная с правил $L_1, \cdots, L_n$ соответственно для каждой ветки, стала не увеличиваться. Те части веток дерева D, начиная от корня, где вероятность возрастала совпадают (это следует из построения дерева семантического вероятностного вывода и существования $\theta_1, \cdots, \theta_n$ подстановок) с частями веток, начиная от корня, дерева семантического вероятностного вывода Tr построенного для отыскания мажорантного наилучшего для предсказания G правила. Пусть совпадение произошло с частями веток $B_1, \cdots, B_n$ дерева Tr. Таким образом правила $P_1, \ldots, P_n$ лежащие в листьях веток $B_1, \cdots, B_n$ имеют вероятности $p(P_i) \geq p(L_i)$. А поскольку η — вероятность мажорантного наилучшего для предсказания G правила не ниже $p(P_i)$, и $p(L_i)) \geq w$ то утверждение теоремы доказано.

Если R не пусто, то существуют $J_1, \cdots, J_s$ — ветки дерева D (их длина $n+1$) такие, что при движении от корня к листьям вероятности правил помечающих вершины встречающиеся на пути возрастает. Пусть $L_1, \cdots, L_s$ — листья этих веток. Тогда в дереве семантического вероятностного вывода есть ветки $B_1, \cdots, B_s$, которые содержат как начальные сегменты ветки $J_1, \cdots, J_s$ дерева D. Пусть опять P_i лежит в листе ветки B_i. Длина веток J_k не превышает длины веток дерева семантического вероятностного вывода, а следовательно для любого $v \in \{1, \cdots, s\}$ имеем $p(P_i) \geq p(L_i)$. С учетом $\eta \geq p(P_i)$ и $p(L_i) = w$ получаем то, что требовалось доказать: $\eta \geq w$. □

Подведем итог. В семантическом подходе

1. Закономерность определяется, как статистический закон. Выделяются особые закономерности:
   - вероятностная закономерность,
   - закономерность, удовлетворяющая RMP,

- закономерность, являющаяся логическим аналогом наилучшего для предсказания правила,
- закономерность, являющаяся логическим аналогом мажорантного наилучшего для предсказания правила.

2. Способ извлечения закономерностей из данных определяется, как семантический вероятностный вывод.
3. Предсказание новых фактов по закономерностям определяется, как индуктивный вывод, основанный на законе, удовлетворяющем RMP и фактах, свойства которых являются посылками статистического закона.

Проиллюстрируем, как при помощи изложенного подхода решается вопрос о возможном быстром выздоровлении Джона Джонса. Пусть статистика людей болевших стрептококком (всего 100 человек) имеет следующее графическое представление, представленное на рис. 4.

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

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

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

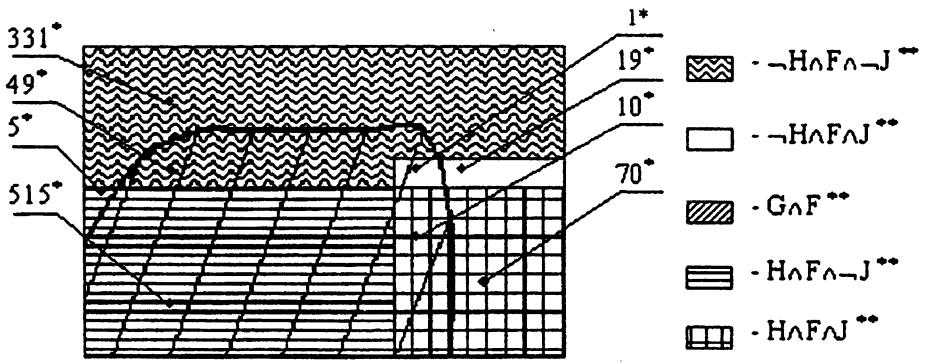
$C_3 : (G(s) = x) \longleftarrow (F(s) = 1), (H(s) = 1), (J(s) = 1); \mu(C_3) = = min\{\mu(C_0), \mu(C_1)\} = 1/8,$

...

где s — переменная для больного стрептококком, интерпретация которой есть конкретный человек из рассматриваемой тысячи, x — переменная, имеющая интерпретацию 0, 1.

Множество фактов состоит из: $(F(b) = 1) \longleftarrow, (H(b)) = = 1) \longleftarrow, (J(b) = 1) \longleftarrow$, их меры — единицы.

Построив множество вероятностных закономерностей, легко убедиться, что для запроса $(G(b) = 1)$ ('Выздоровеет ли быстро Джон Джонс?') мажорантное наилучшее для предсказания правило будет $C^{'} : (G(s) = 1) \longleftarrow (F(s) = 1), (H(s) = 1); \mu(C^{'}) = 0.875$, т.е. ответ — да с вероятностью 0.875. Для запроса $(G(b) = 0)$ (('Верно ли, что Джон Джонс не выздоровеет



* – количество людей в группе,

** – H — пациент лечен, F — болен стрептококком,
J — болен лекарство-устойчивым стрептококком,
G — быстро выздоровел.

Рис. 4. Графическое представление статистической информации о заболевании стрептококком

быстро?')) имеем C'' : $(G(s) = 0) \longleftarrow (F(s) = 1), (J(s) = 1)$; $\mu(C'') = 0.89$, т.е. ответ — нет с вероятностью 0.89. Для запроса $(G(b) = x)$ ('Что произойдет с Джоном Джонсом?') имеем $C : (G(s) = 0) \longleftarrow (F(s) = 1),\ (J(s) = 1); \mu(C) = 0.89$, т.е. ответ — быстрого выздоровления не будет с вероятностью 0.89. Приведенные результаты предсказаний можно получить, построив дерево семантического вероятностного вывода.

Заключение

Определенное в работе семантическое предсказание решает проблемы, возникшие в ранее предложенных подходах. Суть проблем в следующем: модель предсказания Поппера является основой "аксиоматического" подхода к знаниям: "извлечь" из эксперта и поместить в базу знаний (теорию) основополагающие знания (аксиомы, универсальные законы), так чтобы остальные знания и решения получались из них дедуктивным выводом. Проблема такого подхода в том, что для предсказаний требуются тождественно истинные законы (примером являются законы физики), которые могут быть еще не открыты либо не существуют в принципе. Таким образом, из–за наложенного ограничения модель применима лишь для узкого класса задач. То есть задач, где мы априорно знаем свойства всех объектов из определенного класса, которые существовали, существуют или когда–либо возникнут. От нас же требуется лишь дедуктивно предсказать: какое свойство будет (или объяснить наблюдаемое свойство) у конкретного объекта, являющегося частным случаем рассматриваемого класса. В тоже время, в рамках направления Machine Learning и Knowledge Discovery in Data Bases теории не удовлетворяют требованиям Поппера, поскольку являются индуктивными. В этом случае нам известны свойства лишь ограниченного набора объектов из данного класса, а от нас требуется индуктивно предположить (т.е. обобщить имеющиеся данные): какое свойство будет у нового, еще не отраженного в статистической информации объекта в будущем, либо объяснить: почему конкретное свойство фиксируется у него в настоящий момент или наблюдалось в прошлом.

Модель индуктивного вывода Гемпеля решила проблему аксиоматического подхода Поппера, но, как уже отмечалось в ранее (смотри § 2), имеет свои недостатки.

Гемпель предложил RMS, поскольку возникла проблема индуктивной двусмысленности, и в рамках предложенной формализации не было методов её устранения. Но возникает вопрос: насколько эффективно RMS в решении поставленной задачи?

Ограничение однородности слишком сильно чтобы быть проверенным в условиях реально решаемых задач [6], 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. В рассмотренном примере с заболеванием Джона Джонса стрептококком, нам известно, что большинство бацилл гибнут от примененного лекарства, но мы не знаем: применим ли статистический закон из (3) к любому подклассу бацилл или ко всем им в отдельности. Поэтому мы не в праве пользоваться выводом (3) для предсказания быстрого выздоровления Джона Джонса.

Введенное правило RMP позволяет решить проблему индуктивной двусмысленности, поскольку определено: какие индуктивные выводы мы считаем пригодными для предсказания, а какие нет и, исходя из априорных данных, легко проверяется выполнимость условия для конкретного вывода. Сама проверка осуществляется по алгоритму семантического вероятностного вывода мажорантного наилучшего для предсказаний правила.

Более того, было показано, что семантическое предсказание имеет оценку лучшую, чем любое другое индуктивное предсказания при одних и тех же начальных условиях.

Автор выражает благодарность своим научным руководителям Гончарову С.С. и Витяеву Е.Е. за постановку задачи и помощь в проведенном исследовании. Данная работа поддержана грантом: НШ-2112.2003.1.

Литература

1. ПОППЕР К.Р. Логика и рост научного знания. – М.: Прогресс, 1983.

2. ПОППЕР К.Р. Объективное знание, эволюционный подход. – М.: УРСС, 2002.

3. ВИТЯЕВ Е.Е. Семантический подход к созданию баз знаний // Логика и семантическое программирование. — Новосибирск, 1992. — Вып. 146: Вычислительные системы. — С. 19–49.

4. CARNAP R. Logical foundations of probability. - Chicago: University of Chicago Press, 1950.

5. HEMPEL C.G. Deductive–Nomological vs. Statistical Explanation. // Minnesota Studies in the Philosophy of Science III, University of Minnesota Press, Minneapolis, 1962.

6. TARCISIO H.C. PEQUENO, R.S. SILVESTRE. The logic of Intra Theoretical Scientific Reasoning PROC. Of the INT // Workshop on Computational models of scientific reasoning and applications (CMSRA'02).

7. ЗАГОРУЙКО Н.Г. Эмпирическое предсказание. - Новосибирск: Наука, 1979.

8.SHAPIRO E. Algorithmic Program Debugging // MIT Press, 1983, p. 204

9.MATTHEW M. Huntbach An improved vershion of Shapiro's Model Inference system // Third International conference on Logic Programming (Lecture Notes in Computer Science v.225). — P. 180–187.

10. ХАРЛАМОВ Е.Ю., ВИТЯЕВ Е.Е. Формализация понятия предсказание //Вероятностные методы в науке и философии: Материалы региональной конференции. — Новосибирск. Институт философии и права СО РАН. 2003 г. — Новосибирск, 2003. — С. 79-82.

Поступила в редакцию
9 апреля 2004 года