УДК 519.7

Л.В. Шабанова-Кушнаренко

Харьковский национальный университет радиоэлектроники, Харьков

МЕТОД ТАБУЛИРОВАНИЯ МЕТРИКИ ПРЕЦЕДЕНТОВ С ФУНКЦИОНАЛЬНОЙ ЗАВИСИМОСТЬЮ ВЕСОВ ОТ АТРИБУТОВ ПРЕЦЕДЕНТОВ

Общая схема работы информационных систем вывода знаний, основанных на прецедентах (Case-Based Reasoning), состоит из трех действий – поиск похожего прецедента, его адаптация к условиям новой задачи и сохранение в базе знаний нового прецедента. Первое действие делится на три подзадачи - выбор типовых свойств прецедентов, которые будут учитываться при сравнении с аналогичными свойствами новой задачи; собственно сравнение (сопоставление) прецедентов; выбор решения прецедента. Вторая подзадача - сравнения прецедентов, обычно решается с помощью мер (метрик) подобия. Настоящая статья посвящена разработке метода построения метрики на множестве прецедентов с функциональной зависимостью весов от атрибутов прецедентов для оценки их близости.

Ключевые слова: *вывод знаний, прецедент, метрика, вес атрибута прецедента, предикат, Case-Based Reasoning.*

Введение

На сегодняшний день сформировались два основных направления в развитии логического вывода знаний. Первый основан на правилах, второй - на прецедентах.

Вывод, основанный на правилах, требует наличия хорошо формализованной задачи и метода, позволяющего получить ее решение. Однако на практике существует много задач, слабо формализованных, пока не имеющих решения и задач, для которых вообще невозможно найти формальное решение.

При создании экспертной системы, основанной на правилах, необходима трудоемкая работа с экспертными знаниями, затем проверка полученной системы правил на корректность и полноту. Если сравнивать способы работы эксперта по решению задачи с выводом, основанным на правилах, оказывается, что они сильно отличаются – эксперт обычно использует прошлый опыт решений подобных задач и затем, анализируя степень сходства, в зависимости от особенностей конкретной задачи, использует эти решения без существенных изменений или адаптирует их с учетом отличий текущей задачи.

Исследования реальных способов решения задач экспертами привело к созданию метода получения решения, основанного на прецедентах (Case-Based Reasoning, CBR) [1-3]. Этот метод, в отличие от логического вывода, основан на поиске и анализе случаев решения задач, подобных заданной. Поскольку маловероятно, что заданная задача в точности повторит уже решенную и сохраненную в памяти другую задачу, то метод, основанный на прецедентах, обычно требует адаптации найденных решений к новым свойствам и условиям выполнения заданной задачи.

Процесс решения каждой задачи в системе CBR состоит из трех действий – поиск похожего прецедента, его адаптация к условиям новой задачи и сохранение в базе знаний нового прецедента [4].

Обобщение прецедентов влечет за собой существенное усложнение вычислений на этапе подбора ближайшего прецедента, поскольку поиск по базе знаний ведется не по полным отдельным описаниям прецедента, а по факторизованным классам структурных частей множества подобных прецедентов. Обычно в качестве таких частей принимаются подзадачи выбора типовых свойств прецедентов, которые будут учитываться при сравнении с аналогичными свойствами новой задачи; собственно сравнение (сопоставление) прецедентов; выбор решения прецедента [5].

В качестве первой подзадачи обычно используется три основных свойства – цель прецедента, его начальное состояние и сложности, возникающие при решении задачи.

Вторая подзадача - сравнения прецедентов, обычно решается с помощью мер (метрик) подобия – функций, вычисляющих количественное сходство прецедента из базы знаний и новой задачи. В зависимости от сложности предметной области, меры подобия также бывают разной сложности. Самые простые сравнивают число общих подцелей. В более сложных случаях подцелям присваиваются веса и по метрике вычисляется взвешенная сумма общих подцелей. Веса могут быть одинаковыми в пределах каждого класса подобных прецедентов или уникальными для каждого прецедента.

Если сравнение прецедентов выполняется с помощью метрики, то она дает количественную меру их сходства, которая автоматически сортирует прецеденты и, следовательно, решает третью подзадачу выбора решения прецедента.

Использование обобщения усложняет вычисления на третьем этапе - сохранения новых решений по причинам, аналогичным сложностям на этапе описания задачи. Здесь приходится проводить обратные действия по разбиению решения задачи на отдельные, семантически самостоятельные подзадачи и их классификации в соответствии со сформированной в базе знаний структурой классов решенных подзадач.

В статье рассматривается вторая подзадача этапа поиска похожего прецедента - сравнение прецедентов. Разрабатывается метод построения метрики на множестве прецедентов с функциональной зависимостью весов от атрибутов прецедентов.

1. Постановка задачи

Оценка близости прецедентов основана на базовой парадигме CBR: две проблемы близки в том случае, если они имеют близкие решения. Оценка близости может быть выполнена как с помощью формализации логических взаимосвязей между прецедентами в базе знаний, так и с помощью построения оценочной функции.

В первом случае для оценки используется реляционная модель, позволяющая представить оценку в виде бинарного отношения.

Пусть P и Q - прецеденты, сохраненные в базе знаний. Прецедент P содержит m атрибутов $A = \{a_1, a_2, ..., a_m\}$, а прецедент Q содержит n атрибутов $B = \{b_1, b_2, ..., b_n\}$. Простейшим методом оценки близости прецедентов является сравнение числа общих атрибутов, которые также обычно содержатся в базе знаний.

$$C = A \cap B = \{a_1, a_2, ..., a_m\} \cap \cap \{b_1, b_2, ..., b_n\} = \{c_1, c_2, ..., c_k\}, \quad (1)$$

где k - искомое число общих атрибутов прецедентов P и Q.

Назначая пороговое значение параметра k и сравнивая его с вычисленным значением, можно решать задачу классификации прецедентов по заданной степени близости.

Более сложным и точным методом оценки близости прецедентов является присвоение атрибутам весов и вычисление по метрике взвешенной суммы общих подцелей.

Пусть прецедент P имеет множество атрибутов $A = \{a_1, a_2, ..., a_m\}$ с весами $\{v_1, v_2, ..., v_m\}$, тогда оценка P имеет вид

$$s^P = \{v_1 \cdot a_1, v_2 \cdot a_2, ..., v_m \cdot a_m\}.$$

Прецедент Q имеет множество атрибутов $B = \{b_1, b_2, ..., b_n\}$ с весами $\{w_1, w_2, ..., w_n\}$ и оценку $s^Q = \{w_1 \cdot a_1, \ w_2 \cdot a_2, ..., w_n \cdot a_n\}$. Тогда, используя найденные по (1) общие атрибуты $\{c_1, c_2, ..., c_k\}$, берем их оценку с весами $\{t_1, t_2, ..., t_k\}$ и делим на сумму оценок прецедентов P и Q, полученную в результате объединения множеств атрибутов A и B:

$$S = \frac{t_1 \cdot c_1 + t_2 \cdot c_2 + ... + t_k \cdot c_k}{v_1 \cdot a_1, v_2 \cdot a_2, ..., v_m \cdot a_i + w_1 \cdot a_1, w_2 \cdot a_2, ..., w_n \cdot a_j}. \quad (2)$$

Первая модель является достаточно общей, поскольку она дает возможность сравнить атрибуты и структуру прецедентов (взаимосвязи между атрибутами) без учета особенностей предметной области. Во втором случае, при построении функции (метрики) оценки близости S, обычно необходимо использовать переменные, отражающие особенности конкретных моделей.

Рассмотрим задачу оценки близости прецедентов в случае, когда веса атрибутов являются функциями от соответствующих им атрибутов, т.е.

$$S = \sum_{i=1}^{m} v_i(a_i) \cdot a_i. \quad (3)$$

При этом функции $\{v_1, v_2, ..., v_m\}$ могут быть и нелинейными, что значительно усложняет их идентификацию.

Насколько нам известно, вопросы разработки методов определения метрики на множестве прецедентов с учетом функциональной зависимости весов от атрибутов прецедентов до сих пор в литературе не рассматривались. Очевидно, что на степень подобия одного прецедента другим из его класса могут по-разному влиять значения некоторых его атрибутов. Например, если у некоторого атрибута есть критические диапазоны, и его значение попадает в такой диапазон, то это может сильно повлиять на общую оценку прецедента.

2. Разработка метода

Рассмотрим метод нахождения указанных выше весовых функций. Введем предикат подобия E прецедентов P и Q

$$E(P,Q) = \begin{cases} 1, \text{если } (P \in K) \wedge (Q \in K) = 1, \\ 0, \text{если } (P \in K) \wedge (Q \in K) = 0, \end{cases} \quad (4)$$

где P и Q представляют собой m-мерные векторы

$$(v_1(a_1) \cdot a_1, v_2(a_2) \cdot a_2, ..., v_m(a_m) \cdot a_m),$$

$$(w_1(b_1) \cdot b_1, \ w_2(b_2) \cdot b_2, ..., w_m(b_m) \cdot b_m),$$

$(a_1, a_2, ..., a_m)$, $(b_1, b_2, ..., b_m)$ - атрибуты прецедентов P и Q, $\{v_1, v_2, ..., v_m\}$, $\{w_1, w_2, ..., w_m\}$ - веса их атрибутов, K - класс подобных прецедентов.

Введем для краткости записи вектор-функции

$$v = (v_1, v_2, ..., v_m),$$

$$w = (w_1, w_2, ..., w_m),$$

$$a = (a_1, a_2, ..., a_m), \ b = (b_1, b_2, ..., b_m).$$

Предикат E можно записать в виде:

$$E(P,Q) = G((v_1(a_1)\cdot a_1, v_2(a_2)\cdot a_2, ..., v_m(a_m)\cdot a_m), (v_1(b_1)\cdot b_1, v_2(b_2)\cdot b_2, ..., v_m(b_m)\cdot b_m)) = G(v(a)\cdot a, v(b)\cdot b), \quad (5)$$

где G - предикат-классификатор. В целях общности модели предполагаем, что

$$(a_1, a_2, ..., a_m),\ \{v_1, v_2, ..., v_m\} \in R^m$$

и

$$S = \sum_{i=1}^{m} v_i(a_i)\cdot a_i,\quad S \in R^1. \quad (6)$$

где S - скалярная количественная мера, позволяющая классифицировать и сортировать прецеденты по степени сходства.

Найдем достаточное условие линеаризации функций v и w. Задача заключается в нахождении вектор-функции $\alpha = (\alpha_1, \alpha_2, ..., \alpha_m)$, такой, что

$$G(v(a)\cdot a, w(b)\cdot b) = G'(\alpha(v(a))\cdot a, \alpha(w(b))\cdot b), \quad (7)$$

где G' - линейный предикат.

Зная вектор-функцию α, можно перейти от нелинейной модели объекта G к линейной G'. Поскольку вектор-функция α находится по точкам из эксперимента, то построим метод ее табулирования.

1. Определение для функций α_i нулей 0_i, верхних и нижних границ ее значений p_i^+ и p_i^- ($i = \overline{1,m}$).

Только в этом пункте требуется использование явных значений вектор-функции α - величин 0_i, p_i^+, p_i^-. Значениями p_i^+ и p_i^- примем соответственно максимальное и минимальное значения функций α_i. В качестве нулей функций α_i можно брать любые их значения, поскольку они определяют только положение системы координат, в которых находятся. С целью упрощения алгоритма табулирования примем $0_i = p_i^-$. Такой выбор 0_i располагает все остальные значения α_i в положительной области.

Величина шага табулирования h зависит от необходимой точности модели; число получаемых экспериментальных точек k можно вычислить по формуле $k = ((p_i^+ - 0_i)/h) + 1$.

Найдем такие величины δ_i^0, δ_i^h, δ_i^+, что

$$\alpha(\delta_1^0, \delta_2^0, ..., \delta_m^0) = (0_1, 0_2, ..., 0_m),$$
$$\alpha(\delta_1^0, ..., \delta_i^h, ..., \delta_m^0) = (0_1, ..., h, ..., 0_m), \quad (8)$$
$$\alpha(\delta_1^0, ..., \delta_i^+, ..., \delta_m^0) = (0_1, ..., p_i^+, ..., 0_m), i = \overline{1,m}.$$

2. Определение величины δ_2^1, равновеликой величине δ_1^1, т.е. такой, что

$$\alpha_2(\delta_2^1) = \alpha_1(\delta_1^1) \quad (9)$$

Для этого воспользуемся условием

$$G'((\delta_1^1, \delta_2^0, \delta_3, ..., \delta_m), (\delta_1^0, \delta_2^1, \delta_3, ..., \delta_m)) = 1. \quad (10)$$

где $\delta_3, ..., \delta_m$ - произвольные фиксированные величины. Действительно, перейдем к предикату G по формуле (7)

$$G((\alpha_1(\delta_1^1), 0_2, \alpha_3(\delta_3), ..., \alpha_m(\delta_m)), (0_1, \alpha_2(\delta_2^1), \alpha_3(\delta_3), ..., \alpha_m(\delta_m))) = 1. \quad (11)$$

Прежде чем воспользоваться представлением (5) предиката G, выполним его упрощение. Без ограничения общности можно считать, что если $a_i \neq 0$, то $a_i \neq 1$, $i = \overline{1,n}$. Действительно, если $a_i \neq 0$, то можно положить $\alpha'_i(v_i(a_i)) = 1/a_i\alpha_i(v_i(a_i))$. Очевидно, что функции α'_i образуют изоморфизм α' предикатов G и G', такой что $\forall a, b \in R^n$

$$G'(v(a), w(b)) = G(\alpha'_1(v(a)), ..., \alpha'_m(v(a)), \alpha'_1(w(b)), ..., \alpha'_m(w(b)),$$

причем вектор-функция α' и предикат G обладают всеми свойствами вектор-функции α и предиката G' соответственно. Значит, будем считать, что $a_i = 1$ или $a_i = 0$, $i = \overline{1,m}$. Пользуясь последним условием и тем, что $a_1 \neq 0$ и $a_2 \neq 0$, перейдем от (11) к следующему равенству:

$$\alpha_1(\delta_1^1) + 0 + \alpha_3(\delta_3) + ... + \alpha_m(\delta_m) = 0 + \alpha_2(\delta_2^1) + \alpha_3(\delta_3) + ... + \alpha_m(\delta_m). \quad (12)$$

Величины, зависящие от $\delta_3 ... \delta_n$, взаимно уничтожаются и остаются не зависящее от них равенство

$$\alpha_1(\delta_1^1) = \alpha_2(\delta_2^1), \quad (13)$$

что и требовалось показать.

3. Табулирование функции α_1.

3.1. Определение величины δ_1^2, такой что

$$\alpha_1(\delta_1^2) = 2\alpha_1(\delta_1^1). \quad (14)$$

Для этого достаточно воспользоваться условием

$$G'((\delta_1^1, \delta_1^2, \delta_3, ..., \delta_n), (\delta_1^2, \delta_2^0, \delta_3, ..., \delta_n)) = 1. \quad (15)$$

Действительно, используя равенство (7) и условие $a_i = 1$ или $a_i \neq 0$, $i = \overline{1,m}$, получим:

$$\alpha_1(\delta_1^1) + \alpha_2(\delta_2^1) = \alpha_1(\delta_1^2). \quad (16)$$

Вместе с равенством (9) это дает требуемое соотношение (14).

3.2. Определение величины δ_1^3, такой что

$$\alpha_1(\delta_1^3) = 3\alpha_1(\delta_1^1). \quad (17)$$

Для этого следует воспользоваться условием

$$G'((\delta_1^2, \delta_2^1, \delta_3, ..., \delta_n), (\delta_1^3, \delta_2^0, \delta_3, ..., \delta_n)) = 1. \quad (18)$$

Доказательство аналогично предыдущему.

Табулирование продолжается до тех пор, пока не будет пройден весь интересующий нас диапазон, т.е. до шага $3.N$, когда

$$\delta_l^N \geq \delta_l^{p^+}, \delta_l^{N-1} < \delta_l^{p^+}. \quad (19)$$

4. Табулирование функций α_i, $i = \overline{2,m}$, таких что $a_i \neq 0$. Для определения величины δ_i^k, такой что

$$\alpha_i(\delta_i^k) = kh \quad (20)$$

следует воспользоваться условием

$$G'((\delta_1^k, \delta_2, ..., \delta_{i-1}, \delta_i^0, \delta_{i+1}, ..., \delta_m), (\delta_1^0, \delta_2, ..., \delta_{i-1}, \delta_i^k, \delta_{i+1}, ..., \delta_m)) = 1, \quad (21)$$

$i = \overline{2,m}$, $k = \overline{1,N}$.

Доказательство такое же, как и в п. 3.1. Если $a_i = 0$, то значения δ_i и α_i не влияют на предикаты G и G'. Общая таблица табулирования имеет вид, представленный в табл. 1.

Таблица 1

Табулирование вектор-функции α

$\alpha_i(v_i(a_i))$	0	h	2h	...	Nh
$v_i(a_i)$	δ_i^0	δ_i^1	δ_i^2	...	δ_i^N

Табулирование закончено.

Выводы

Метод получения знаний на основе вывода по прецедентам сейчас широко используется в системах доступа к информации и автоматизации информационно-справочных служб, в медицинской диагностике, юриспруденции, для мониторинга и диагностики технических систем, при поиске решения в проблемных ситуациях.

Параметры, по которым решенная задача отбирается из базы знаний в качестве наиболее близкого прецедента, уникальны для каждой CBR-системы, поскольку они определяются свойствами класса типичных задач, под который разрабатывалась система. Определение весов атрибутов прецедента – обычно сложный итерационный процесс, выполняемый экспертом. При этом должны учитываться многие параметры, среди которых – приоритеты важности подзадач, решаемых в прецеденте, сложность решения этих подзадач, их независимость друг от друга.

В статье разработан метод табулирования метрики на множестве прецедентов с функциональной зависимостью весов от атрибутов прецедентов, который линеаризует функциональную зависимость на основе изменения шкал измерения атрибутов, что позволяет оценить близость темпоральных моделей в базе прецедентов путем сравнения последовательности временных меток действий или событий.

Для некоторых предметных областей атрибуты прецедентов могут иметь различные веса в зависимости от своих значений (в том числе с нелинейной характеристикой зависимости). Например, если у некоторого атрибута есть критические диапазоны, и его значение попадает в такой диапазон, то это может сильно повлиять на общую оценку прецедента. Такая ситуация характерна для ИСППР реального времени при мониторинге и управлении сложными объектами или процессами.

Список литературы

1. Aamodt, A., Plaza, E. Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. AI Communications, 7(1), 1994. – P. 39-59.

2. Schank Roger. Dynamic Memory: A Theory of Learning in Computers and People / Roger Schank // New York: Cambridge University Press, 1982.

3. Awad E. Knowledge Management / E. Awad, H. Ghaziri. – Prentice Hall, 2004. - 480 p.

4. Hanks S. A Domain-Independent Algorithm for Plan Adaptation / S. Hanks, D.S. Weld // Journal of Artificial Intelligence Research. – 1995. – Vol. 2.– P. 319-360.

5. Peter Funk, Pedro A. González-Calero (Eds.): Advances in Case-Based Reasoning, 7th European Conference, ECCBR 2004, Madrid, Spain, August 30 - September 2, 2004, Proceedings. Lecture Notes in Computer Science 3155 Springer 2004, ISBN 3-540-22882-9 Contents BibTeX - ECCBR 2004 Home Page.

Поступила в редколлегию 30.07.2015

Рецензент: д-р техн. наук, проф. В.А. Филатов, Харьковский национальный университет радиоэлектроники, Харьков.

МЕТОД ТАБУЛЮВАННЯ МЕТРИКИ ПРЕЦЕДЕНТІВ ІЗ ФУНКЦІОНАЛЬНОЮ ЗАЛЕЖНОСТЮ ВАГІВ ВІД АТРИБУТІВ ПРЕЦЕДЕНТІВ

Л.В. Шабанова-Кушнаренко

Розробляється метод побудови метрики на множині прецедентів із функціональною залежністю ваг від атрибутів прецедентів. Така метрика необхідна при вирішенні завдань порівняння прецедентів, їх класифікації та пошуку прецедентів, максимально схожих на задані завдання, в базах знань інформаційних систем, що використовують методи виведення знань, засновані на прецедентах.

Ключові слова: *висновок знань, прецедент, метрика, вага атрибуту прецеденту, предикат, Case-Based Reasoning.*

THE PRECEDENT METRICS TABULATION METHOD WITH FUNCTIONAL WEIGHT DEPENDENCE ON THE PRECEDENT ATTRIBUTES

L.V. Shabanova-Kushnarenko

The developed is a method of constructing a metric on the set of precedents with functional weight dependence on the precedent attributes. This metric is needed for the precedents comparison, their classification and search of precedents that are the most similar to the given tasks in the knowledge bases of information systems using conclusion knowledge methods based on precedents.

Keywords: *conclusion knowledge, precedent, metric, attribute weight, precedent, predicate, Case-Based Reasoning.*