УДК 004.032.2 DOI: 10.30748/soi.2017.149.08

Д.Э. Ситников[1], П.Э. Ситникова[2]

[1] *Харьковская государственная академия культуры, Харьков*

[2] *Харьковский гуманитарный университет «Народная украинская академия», Харьков*

ЛОГИКО-АЛГЕБРАИЧЕСКИЙ ПОДХОД К ПОСТРОЕНИЮ ДЕРЕВЬЕВ РЕШЕНИЙ И ОПЕРАЦИЯМ С НИМИ

В данной работе предложен логико-алгебраический подход к извлечению логических правил из баз данных на основе построения деревьев решений. Проанализированы основные этапы генерации логических правил. Предложенный подход позволяет работать как с реляционными базами данных, так и с базами знаний в виде логических уравнений. В качестве основной операции исключения нецелевых переменных из исходного уравнения рассматривается операция навешивания квантора существования.

Ключевые слова: *деревья решений, CART-алгоритм, ID3-алгоритм, информационная энтропия, индекс Gain*.

Введение

В настоящее время интенсивно развивается научно-практическая область исследований, связанная с интеллектуализацией методов обработки и анализа данных. Это научное направление характеризуется сочетанием широкого математического инструментария (от классического статистического анализа до новых кибернетических методов) и последних достижений в сфере информационных технологий.

Метод деревьев решений (decision trees) является одним из наиболее популярных методов решения задач классификации и прогнозирования. Иногда этот метод Data Mining также называют деревьями решающих правил, деревьями классификации и регрессии. Дерево решений представляет собой граф, в котором из каждой вершины исходит множество альтернатив, а каждый лист дерева – это возможное решение.

Разработан ряд алгоритмов, реализующих этот метод Data Mining. Это алгоритмы CART, ID3 и его расширение C4.5. Соответственно эти алгоритмы решают задачу классификации данных и на этом основании классифицировать новые данные. Например, мы можем записать значения определенных параметров, называемых атрибутами, таких как пол, уровень дохода, возраст и др. и действия данных людей в определенном случае, например, возврат или невозврат кредита. Или же описав с помощью определенных атрибутов состояние здоровья (температура, головная боль, пульс, давление и др.) и соответствующий диагноз, иметь возможность поставить диагноз новым пациентам.

Несомненным преимуществом методов, основанных на деревьях решений, является явное выражение искомых скрытых зависимостей между атрибутами данных. В отличие, например, от применения искусственных нейронных сетей, в дереве решений после формирования итоговой системы правил, связывающих целевую переменную (target variable) с остальными переменными, можно проанализировать всю последовательность переходов в явном виде.

Существующие методы построения деревьев решений имеют дело с реляционными таблицами, что существенно ограничивает возможности интеллектуального анализа данных с более сложной структурой. В данной работе ставится задача описания этапов построения дерева решений в виде последовательности логических уравнений, в которых с помощью кванторных операций исключаются нецелевые переменные. Предполагается, что такой подход позволит использовать в дальнейшем основные идеи существующих методов генерации деревьев для извлечения логических зависимостей из более сложных структур данных и знаний, в частности, из сетевых и иерархических структур.

Целью данной работы является описание процедуры и анализа построения дерева решений на логико-алгебраическом языке.

Логико-алгебраическое описание процесса построения дерева

Опишем построение дерева решений на следующем примере. Предположим, есть некоторый набор данных, касающийся определенной группы людей: *пол*, *владение автомобилем*, *стоимость путешествия*, *доход*, а также выбранный каждым человеком *транспорт для путешествий*.

Составив модель данной задачи, можно предположить, что каждый человек описывается сле-

дующими свойствами (атрибутами): пол – со значениями {мужской, женский}, владение автомобилем – со значениями {0,1,2}, стоимость путешествия – со значениями {дешевый, стандартный, дорогой}, доход – со значениями {низкий, средний, высокий}. Необходимо построить дерево решений, которое бы определило предпочтительный вид транспорта для любого нового человека, например, для имеющего один автомобиль, стандартную стоимость путешествия и высокий уровень дохода. Построив дерево решений, мы сможем пройти по нему в соответствии с нужными значениями атрибутов и дойти до соответствующего листа, содержащего конечное результирующее значение.

Представим данные статистике в виде таблицы. При этом для каждого случая известно значение результирующей функции.

Таблица 1

Данные статистики описываемой задачи

	Пол	Авто	Стоимость путешествия	Доход	Транспорт
1.	Муж	0	Дешевый	Низкий	Автобус
2.	Муж	1	Дешевый	Средний	Автобус
3.	Жен	1	Дешевый	Средний	Поезд
4.	Жен	0	Дешевый	Низкий	Автобус
5.	Муж	1	Дешевый	Средний	Автобус
6.	Муж	0	Стандартный	Средний	Поезд
7.	Жен	1	Стандартный	Средний	Поезд
8.	Жен	1	Дорогой	Высокий	Автомобиль
9.	Муж	2	Дорогой	Средний	Автомобиль
10.	Жен	2	Дорогой	Высокий	Автомобиль

Данные выборки можно также представить с использованием аппарата алгебры конечных предикатов (АКП) [1; 2]. Поставим каждому атрибуту в соответствие переменную с множеством соответствующих значений. Таким образом, предикат, описывающий выборку, будет определен на множестве переменных $\{x_1, x_2, x_3, x_4, x_5\}$, которые имеют соответствующие области определения. Например, атрибут пол (переменная x_1 со значением «мужской (male)» – *m* и «женский (female)» – *f*) обозначим как x_1^m и x_1^f, атрибут количество авто (переменная x_1 со значениями 0, 1, 2) – x_2^0, x_2^1, x_2^2 стоимость путешествия со значениями дешевый (cheap), стандартный (standart), дорогой (expensive) – x_3^c, x_3^s, x_3^e доход со значениями низкий (low) , высокий (high), средний (medium) – x_4^s, x_4^h, x_4^m; вид транспорта со значениями автобус (bus), поезд (train), автомобиль (avto) – x_5^b, x_5^t, x_5^a.

Предикат, соответствующий приведенной выше таблице, будет описан следующим образом:

$$\begin{aligned} P(x_1,x_2,x_3,x_4,x_5) = {} & x_1^m x_2^0 x_3^c x_4^l x_5^b \vee x_1^m x_2^1 x_3^c x_4^m x_5^b \vee \\ & \vee x_1^f x_2^1 x_3^c x_4^m x_5^t \vee x_1^f x_2^0 x_3^c x_4^l x_5^b \vee x_1^m x_2^1 x_3^c x_4^m x_5^b \vee \\ & \vee x_1^m x_2^0 x_3^c x_4^m x_5^t \vee x_1^f x_2^1 x_3^s x_4^m x_5^t \vee x_1^f x_2^1 x_3^e x_4^h x_5^a \vee \\ & \vee x_1^m x_2^2 x_3^e x_4^m x_5^a \vee x_1^f x_2^2 x_3^e x_4^h x_5^a. \end{aligned} \quad (1)$$

Такое представление имеет вид дизъюнктивной нормальной формы (ДНФ) и в общем случае является более компактным, чем табличное представление данных, поскольку пропуски и дублирование данных влекут разрастание реляционной таблицы, в то время как ДНФ-представление будет учитывать пропуски просто добавлением элементарных конъюнкций. Логико-алгебраический способ представления данных дает возможность извлекать логические зависимости из любых структур данных, которые могут быть представлены в виде системы логических уравнений. Для этого можно использовать кванторы существования и общности. В работе [3] описываются структуры, в которых навешивание кванторов не приводит к усложнению формулы, что позволяет эффективно обрабатывать большие массивы информации.

Первым шагом при построении дерева решений является выбор некоторого атрибута, по которому будет проведено разбиение обучающей выборки на классы. Различные алгоритмы предполагают разные подходы в данном случае. Анализируя работу алгоритма ID3, мы сталкиваемся с понятием энтропии.

Энтропия может интерпретироваться как мера неопределённости (неупорядоченности) некоторой системы. Чем больше единообразие, чем больше энтропия, тем больше информации мы можем извлечь.

Информационная энтропия описывается множеством символов $x_1,...,x_n$ и вероятностей $p_1,...,p_n$ появления этих символов в сообщении.

В теории информации энтропия вычисляется по формуле, связанной с дискретным распределением вероятностей

$$E(S) = \sum_{x \in X} -p(x)\log_2 p(x).$$

Информационная энтропия равна нулю, когда какая-либо вероятность равна единице (а остальные – нулю), т. е. когда информация полностью предсказуема и не несёт ничего нового для приёмника. Энтропия принимает наибольшее значение для равновероятного распределения, когда все вероятности

p_k одинаковы; т. е. когда неопределенность, разрешаемая сообщением максимальна.

Если есть множество S, состоящее из n элементов, и есть свойство A, которое может принимать T значений, каждое из которых реализуется в m_t случаях, то энтропия множества S по отношению к свойству A вычисляется по формуле:

$$E(S,A) = -\sum_{t \in T} \frac{m_t}{n} \log_2 \frac{m_t}{n}. \qquad (2)$$

Посчитаем энтропию системы для приведенной задачи. Общее количество наблюдений в нашем примере: $n = 10$, при этом $m_1 = 4$ (количество случаев с выбором автобуса), $m_2 = 3$ (поезда), $m_3 = 3$ (автомобиля). Таким образом, вероятность события выбора каждого вида транспорта $\left(\frac{m_i}{n}\right)$, соответственно 0,4; 0,3 и 0,3 соответственно. Подсчитаем энтропию:

$$\text{Entropy}(S) = -0,4 * \log_2 0,4 - 0,3 * \log_2 0,4 - \\ -0,3 * \log_2 0,3 = 1,571.$$

Вернемся к вопросу о том, какой атрибут выбрать для первоначальной классификации (который будет в корне дерева решений)? Из понятия энтропии следует, что атрибут должен быть таким, чтобы после классификации энтропия стала как можно меньше. Здесь мы сталкиваемся с понятием «прироста информации» или индексом *Information Gain*.

Индекс *Gain* – это мера разницы начальной энтропии и энтропии, полученной после деления по данному атрибуту.

Предположим, есть множество S элементов, которое классифицировано по атрибуту A. S_t – множество элементов в подмножестве, на котором атрибут A принимает значение t.

Подсчитываем индекс Gain по формуле:

$$\text{Gain}(S,A) = E(S) - \sum_{t \in T} p(t) E(S_t),$$

где

– $E(S)$ – энтропия множества S;

– A – атрибут, по которому проведена классификация;

– T – подмножество, образованное при расщеплении множества S по атрибуту A такое, что $S = \bigcup_{t \in T} t$;

– $p(t)$ – отношение количества элементов в подмножестве S_t к общему количеству элементов в S;

– $E(S,S_t)$ – энтропия множества S по отношению к подмножеству S_t.

Вычислим индекс *information Gain* для каждого атрибута.

Рассмотрим атрибут *пол – мужской (m)* и *женский (f)*. Деление по классам для этого атрибута следующее (табл. 2).

Таким образом, в подмножествах пол^m и пол^f образуется по 5 элементов:

$$\left|S_{\text{пол}^m}\right| = 5, \ \left|S_{\text{пол}^f}\right| = 5, \ |S| = 10.$$

Таким образом, для значения пол^m в этом частном случае $n = 5$, при этом $m_1 = 3$ (количество случаев с выбором автобуса), $m_2 = 1$ (поезда), $m_3 = 1$ (автомобиля).

Таблица 2

Деление множества по атрибуту «пол»

Пол	Транспорт	Пол	Транспорт
Муж	Автобус	Жен	Поезд
Муж	Автобус	Жен	Автобус
Муж	Автобус	Жен	Поезд
Муж	Поезд	Жен	Автомобиль
Муж	Автомобиль	Жен	Автомобиль

И вероятности, соответственно: автобус – 3/5, поезд – 1/5, автомобиль – 1/5. Энтропия в этом случае:

$$E(S, \text{пол}^m) = 0,6 * \log_2 0,6 - 0,2 * \log_2 0,2 - \\ -0,2 * \log_2 0,2 = 1,522.$$

Для значения пол^f имеем: $n = 5$, при этом $m_1 = 1$ (количество случаев с выбором автобуса), $m_2 = 2$ (поезда), $m_3 = 2$ (автомобиля).

Для значения пол^f подсчитаем вероятности: автобус – 1/5, поезд – 2/5, автомобиль – 2/5. Энтропия в этом случае:

$$E(S, \text{пол}^f) = 0,2 * \log_2 0,2 - 0,4 * \log_2 0,4 - \\ -0,4 * \log_2 0,4 = 1,371.$$

Тогда

$$\text{Gain}(S, \text{пол}) = 1,571 - ((5/10) * 1,522 + (5/10) * \\ * 1,371) = 0,12.$$

На языке АКП мы можем сформулировать вопрос следующим образом: какова связь между атрибутом «пол» и атрибутом «транспорт»? Для установления этой связи следует исключить остальные переменные с помощью квантора существования:

$$\exists x_2, x_3, x_4 (P(x_1, x_2, x_3, x_4, x_5)) = x_1^m x_5^b \vee x_1^m x_5^b \vee \\ \vee x_1^f x_5^t \vee x_1^f x_5^b \vee x_1^m x_5^b \vee x_1^m x_5^t \vee x_1^f x_5^t \vee x_1^f x_5^a \vee \\ \vee x_1^m x_5^a \vee x_1^f x_5^a = x_1^m x_5^b \vee x_1^f x_5^t \vee x_1^f x_5^b \vee x_1^m x_5^t \vee \\ \vee x_1^f x_5^a \vee x_1^m x_5^a = x_1^m (x_5^b \vee x_5^t \vee x_5^a) \vee \\ \vee x_1^f (x_5^t \vee x_5^b \vee x_5^a) = x_1^m \vee x_1^f = 1.$$

Полученное тождественно истинное высказывание говорит о том, что без учета атрибутов x_2, x_3, x_4 между атрибутами «пол» и «транспорт» нет никакой логической связи. Эти атрибуты могут принимать любые значения из своих областей определения независимо друг от друга. Следовательно, можно утверждать, что без промежуточных переменных значение атрибута «пол» не влияет на значения целевой переменной «транспорт». Очевидно, что результат соответствует данным, приведенным в табл. 1.

Рассмотрим атрибут *владение автомобилем (auto)*. Деление по классам для этого атрибута следующее (табл. 3).

Таблица 3

Деление множества по атрибуту «владение автомобилем»

Авто	Транспорт	Авто	Транспорт	Авто	Транспорт
0	Автобус	1	Автобус	2	Автом.
0	Автобус	1	Поезд	2	Автом.
0	Поезд	1	Автобус		
		1	Поезд		
		1	Автом.		

Легко подсчитать энтропию и индекс *Gain* для разбиения по данному атрибуту:

$$E(S, avto^0) = -2/3 * \log_2(2/3) - 1/3 * \log_2(1/3) - -0 * \log_2 0 = 0,918.$$

$$E(S, avto^1) = -2/5 * \log_2(2/5) - 2/5 * \log_2(2/5) - -1/5 * \log_2 1/5 = 1,522.$$

$$E(S, avto^2) = -2/2 * \log_2(2/2) - 0 * \log_2 0 - -0 * \log_2 0 = 0.$$

$$Gain(S, avto) = 1,571 - ((3/10 * 0,918 + +(5/10) * 1,522 + (2/10) * 0) = 0,534.$$

Соответственно, для установления связи между признаками *владение автомобилем* и *транспорт* запишем следующее исключение переменных:

$$\exists x_1, x_3, x_4 (P(x_1, x_2, x_3, x_4, x_5)) = x_2^0 x_5^b \vee x_2^1 x_5^b \vee \vee x_2^1 x_5^t \vee x_2^0 x_5^b \vee x_2^1 x_5^b \vee x_2^0 x_5^t \vee x_2^1 x_5^t \vee x_2^1 x_5^a \vee x_2^2 x_5^a \vee \vee x_2^2 x_5^b = x_2^0 x_5^b \vee x_2^1 x_5^b \vee x_2^1 x_5^t \vee x_2^0 x_5^t \vee x_2^1 x_5^a \vee x_2^2 x_5^a = = x_2^0 (x_5^b \vee x_5^t) \vee x_2^1 (x_5^b \vee x_5^t \vee x_5^a) \vee x_5^a x_2^2 = = x_2^0 (x_5^b \vee x_5^t) \vee x_2^1 \vee x_5^a x_2^2.$$

Полученное выражение можно записать в виде импликаций:

$$x_2^0 \to (x_5^b \vee x_5^t),\ x_2^2 \to x_5^a.$$

В данном случае можно утверждать, что владение двумя автомобилями позволяет однозначно классифицировать транспорт (автомобиль), отсутствие автомобиля влечет выбор поезда или автобуса, а владение единственным автомобилем не позволяет ничего сказать о целевом атрибуте без привлечения дополнительных атрибутов.

Рассмотрим атрибут *стоимость путешествия (trav)*. Деление по классам для этого атрибута следующее (табл. 4).

Таблица 4

Деление по атрибуту «стоимость путешествия»

Стоим. путеш.	Транспорт	Стоим. путеш.	Транспорт	Стоим. путеш.	Транспорт
Дешевый	Автобус	Станд.	Поезд	Дорогой	Авто
Дешевый	Автобус	Станд.	Поезд	Дорогой	Авто
Дешевый	Поезд			Дорогой	Авто
Дешевый	Автобус				
Дешевый	Автобус				

Подсчитаем индекс *Gain:*

$$E(S, trav^{cheap}) = -4/5 * \log_2(-4/5) - 1/5 * \log_2(1/5) - -0 * \log_2 0 = 0,722.$$

$$E(S, trav^{stand}) = -2/2 * \log_2(2/2) = 0.$$

$$E(S, trav^{exp}) = -3/3 * \log_2(3/3) = 0.$$

$$Gain(S, trav) = 1.571 - ((5/10) * 0,722 + +(2/10) * 0 + (3/10) * 0) = 1,21.$$

На языке АКП связь между атрибутами *стоимость путешествия* и *транспорт* будет следующей:

$$\exists x_1, x_2, x_4 (P(x_1, x_2, x_3, x_4, x_5)) = x_3^c x_5^b \vee x_3^c x_5^b \vee \vee x_3^c x_5^t \vee x_3^c x_5^b \vee x_3^c x_5^b \vee x_3^c x_5^t \vee x_3^s x_5^t \vee x_3^e x_5^a \vee \vee x_3^e x_5^a \vee x_3^e x_5^a = x_3^c x_5^b \vee x_3^c x_5^t \vee x_3^s x_5^t \vee x_3^e x_5^a = = x_3^c (x_5^b \vee x_5^t) \vee x_3^s x_5^t \vee x_3^e x_5^a.$$

Полученное выражение можно записать в виде следующей системы правил:

$$x_3^c \to (x_5^b \vee x_5^t),\ x_3^s \to x_5^t,\ x_3^e \to x_5^a.$$

Из найденных импликаций видно, что выбор и стандартного, и дорогого путешествия позволяет однозначно определить транспорт, а выбор дешевого путешествия оставляет некоторую неопределенность относительно транспорта. Заметим, однако, что в этом случае можно однозначно утверждать, что автомобиль не будет выбран.

Рассмотрим атрибут *доход*. Деление по классам для этого атрибута следующее (табл. 5).

В данном случае имеем:

$$E(S, income^m) = -2/6 * \log_2(2/6) - 1/6 * \log_2(1/6) - -3/6 * \log_2(3/6) = 1,459.$$

$$\text{Gain}(S,\text{income}) = 1,571 - ((2/10)*0 + (2/10)*0 + \\ +(6/10)*1,459 = 0,695.$$

Таблица 5

Деление по атрибуту «доход»

До-ход	Транс-порт	До-ход	Тран спор т	Доход	Транс порт
Вы-сокий	Авто-мобиль	Низ-кий	Авто бус	Сред-ний	Авто-бус
Вы-сокий	Авто-мобиль	Низ-кий	Авто бус	Сред-ний	Поезд
				Сред-ний	Авто-бус
				Сред-ний	Поезд
				Сред-ний	Поезд
				Сред-ний	Авто-мо-биль

С помощью уравнения АКП представление связи между атрибутами *доход* и *транспорт* будет следующей:

$$\exists x_1,x_2,x_3(P(x_1,x_2,x_3,x_4,x_5)) = x_4^l x_5^b \vee x_4^m x_5^b \vee \\ \vee x_4^m x_5^t \vee x_4^l x_5^b \vee x_4^m x_5^b \vee x_4^m x_5^t \vee x_4^m x_5^t \vee x_4^h x_5^a \vee \\ \vee x_4^m x_5^a \vee x_4^h x_5^b = x_4^l x_5^b \vee x_4^m x_5^b \vee x_4^m x_5^t \vee x_4^h x_5^a \vee x_4^m x_5^a = \\ = x_4^l x_5^b \vee x_4^m (x_5^b \vee x_5^t \vee x_5^a) \vee x_4^h x_5^a = x_4^l x_5^b \vee x_4^m \vee x_4^h x_5^a.$$

или, в виде импликаций: $x_4^l \to x_5^b$, $x_4^h \to x_5^a$.

В этом случае низкий и высокий доход позволяют однозначно определить транспорт, тогда как средний доход не позволяет ничего сказать о выборе транспортного средства.

Выпишем индексы *Information Gain* для различных атрибутов:

Атрибут	*Information Gain*
Пол	0,12
Авто	0,534
Стоимость путешествия	1,21
Доход	0,695

Максимальный индекс *Information Gain* для атрибута *Стоимость путешествия*. Анализируя полученные системы импликаций для каждого атрибута, можно заметить, что не только с точки зрения *Information Gain*, но и с логической точки зрения стоимость путешествия – самый информативный признак для определения транспортного средства, так как два значения этого признака позволяют однозначно выбрать транспорт, а третье значение сужает возможность выбора до двух видов транспорта.

Поместив данный атрибут в вершине дерева, получаем первый уровень расщепления (рис. 1).

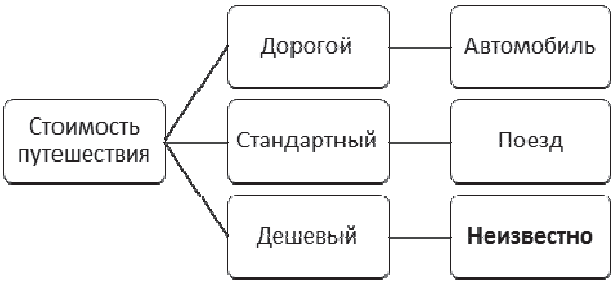

Рис. 1. Дерево решений для первого уровня расщепления

Таким образом, можно сказать, что если стоимость путешествия принимает значение «дорогой», то это определенно влечет выбор транспорта – «автомобиль».

То есть, применяя оператор подстановки к исходному предикату, мы можем выразить данную зависимость уравнением:

$$x_3^e(P) \to x_5^a.$$

Действительно,

$$x_3^e(P(x_1,x_2,x_3,x_4,x_5)) = P(x_1,x_2,e,x_4,x_5) = \\ = x_1^f x_2^l x_4^h x_5^a \vee x_1^m x_2^2 x_4^m x_5^a \vee x_1^f x_2^2 x_4^h x_5^b.$$

и

$$\exists x_1,x_2,x_4(P(x_1,x_2,e,x_4,x_5)) = x_5^a \vee x_5^a \vee x_5^a = x_5^a.$$

Аналогично, стоимость путешествия со значением «стандартная» определенно влечет выбор в качестве транспорта «поезд». Запишем с использованием подстановочного оператора:

$$x_3^s(P) \to x_5^t.$$

$$x_3^s(P(x_1,x_2,x_3,x_4,x_5)) = P(x_1,x_2,s,x_4,x_5) = \\ = x_1^m x_2^0 x_4^m x_5^t \vee x_1^f x_2^l x_4^m x_5^t;$$

$$\exists x_1,x_2,x_4(P(x_1,x_2,s,x_4,x_5)) = x_5^t \vee x_5^t = x_5^t.$$

Вместе с тем, применяя оператор подстановки $x_3^c(P)$, то есть, выбрав «стандартную» стоимость путешествия, получим

$$x_3^c(P(x_1,x_2,x_3,x_4,x_5)) = P(x_1,x_2,c,x_4,x_5) = \\ = x_1^m x_2^0 x_4^l x_5^b \vee x_1^m x_2^l x_4^m x_5^b \vee x_1^f x_2^l x_4^m x_5^t \vee \\ \vee x_1^f x_2^0 x_4^l x_5^b \vee x_1^m x_2^l x_4^m x_5^b.$$

Соответственно,

$$\exists x_1,x_2,x_4(P(x_1,x_2,c,x_4,x_5)) = x_5^b \vee x_5^b \vee x_5^t \vee x_5^b \vee \\ \vee x_5^b = x_5^b \vee x_5^t.$$

Теперь необходимо повторить описанный алгоритм для следующей вершины, поскольку дерево еще не построено.

Выберем данные, для атрибута *способ путешествия* со значением *дешевый* (Табл. 6).

Таблица 6

Данные для атрибута «способ путешествия» со значением «дешевый»

Пол	Авто	Доход	Транспорт
Мужской	0	Низкий	Автобус
Женский	0	Низкий	Автобус
Мужской	1	Средний	Автобус
Женский	1	Средний	Поезд
Мужской	1	Средний	Автобус

Общее количество наблюдений на данном этапе: $n = 5$, при этом $m_1 = 4$ (количество случаев с выбором автобуса), $m_2 = 1$ (поезда).

Соответственно можем вычислить энтропию данной системы:

$$E(S) = -0,8 * \log_2 0,8 - 0,2 * \log_2 0,2 = 0,722.$$

После того, как стал известен признак, который был помещен в вершину дерева (стоимость путешествия), и определения ветки, по которой надо найти разбиение, уравнение (1) будет выглядеть так:

$$P(x_1, x_2, c, x_4, x_5) = x_1^m x_2^0 x_4^l x_5^b \vee x_1^m x_2^l x_4^m x_5^b \vee$$
$$\vee x_1^f x_2^l x_4^m x_5^t \vee x_1^f x_2^0 x_4^l x_5^b \vee x_1^m x_2^l x_4^m x_5^b.$$

Теперь опять подсчитаем *Information Gain* для каждого атрибута. Рассмотрим атрибут *пол*.

Таблица 7

Деление по атрибуту «пол»

Пол	Транспорт	Пол	Транспорт
Женский	Автобус	Мужской	Автобус
Женский	Поезд	Мужской	Автобус
		Мужской	Автобус

Очевидно, для атрибута пол^m энтропия будет равна нулю, поскольку вероятность для значения пол$^m = 3/3 = 1$.

Для атрибута пол^f получаем: вероятности по ½ для автобуса и поезда. Таким образом,

$$\text{Entropy}(S) = -0,5 * \log_2 0,5 - 0,5 * \log_2 0,5 = 1.$$

Подсчитаем индекс Gain:

$$\text{Gain}(S_2, s) = 0,722 - ((3/5) * 0 + (2/5) * 1 = 0,322.$$

При этом уравнение принимает вид

$$\exists x_2, x_4 (P(x_1, x_2, c, x_4, x_5)) = x_1^m x_5^b \vee x_1^m x_5^b \vee x_1^f x_5^t \vee$$
$$\vee x_1^f x_5^b \vee x_1^m x_5^b = x_1^m x_5^b \vee x_1^f x_5^t \vee x_1^f x_5^b = x_1^m x_5^b \vee$$
$$\vee x_1^f (x_5^t \vee x_5^b).$$

Аналогично, рассмотрим атрибут *Количество автомобилей*.

Таблица 8

Расщепление по атрибуту «Количество авто»

Авто	Транспорт	Авто	Транспорт
0	Автобус	1	Поезд
0	Автобус	1	Автобус
		1	Автобус

Соответственно, для атрибута avto^0 энтропия будет равна нулю.

Для атрибута avto^1 вероятности равны 1/3 (поезд) и 2/3 (автобус).

$$\text{Entropy}(S, \text{avto}^1) = -1/3 * \log_2(1/3) - 2/3 * \log_2(2/3) =$$
$$= 0,918.$$

Соответственно индекс Gain:

$$\text{Gain}(S_2, \text{avto}) =$$
$$= 0,722 - ((2/5) * 0 + (3/5) * 0,918 = 0,171.$$

При этом уравнение принимает вид

$$\exists x_1, x_4 (P(x_1, x_2, c, x_4, x_5)) = x_2^0 x_5^b \vee x_2^l x_5^b \vee x_2^l x_5^t \vee$$
$$\vee x_2^0 x_5^b \vee x_2^l x_5^b = x_2^0 x_5^b \vee x_2^l x_5^b \vee x_2^l x_5^t =$$
$$= x_2^0 x_5^b \vee x_2^l (x_5^b \vee x_5^t).$$

Аналогично, рассмотрим атрибут *Доход*.

Таблица 9

Расщепление по атрибуту «доход»

Доход	Транспорт	Доход	Транспорт
Низкий	Автобус	Средний	Автобус
Низкий	Автобус	Средний	Поезд
		Средний	Автобус

Для атрибута income^{low} энтропия будет равна нулю. Для атрибута income^{middle}:

$$E(S, \text{income}^{high}) = -1/3 * \log_2(1/3) -$$
$$-2/3 * \log_2(2/3) = 0,918.$$
$$\text{Gain}(S, \text{income}) = 0,722 - ((2/5) * 0 +$$
$$+(3/5) * 0,918) = 0,171.$$

При этом уравнение принимает вид

$$\exists x_1, x_2 (P(x_1, x_2, c, x_4, x_5)) = x_4^l x_5^b \vee x_4^m x_5^b \vee$$
$$\vee x_4^m x_5^t \vee x_4^l x_5^b \vee x_4^m x_5^b = x_4^l x_5^b \vee x_4^m x_5^b \vee x_4^m x_5^t =$$
$$= x_4^l x_5^b \vee x_4^m (x_5^b \vee x_5^t).$$

Заметим, что все предикаты, полученные в результате исключения переменных, имеют более простую и легкую для интерпретации структуру.

Выпишем индексы *Information Gain* для различных атрибутов:

Атрибут	*Information Gain*
Пол	0,322
Авто	0,171
Доход	0,171

Поскольку индекс *Information Gain* для атрибута *пол* является максимальным, проводим следующее расщепление по этому атрибуту. Таким образом, можем построить дерево решений (рис. 2).

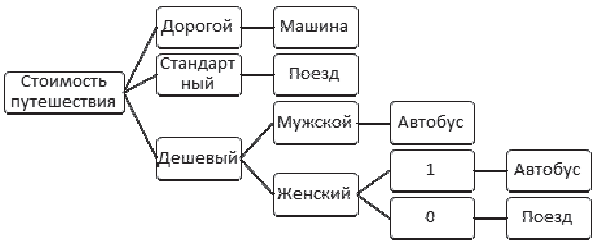

Рис. 2. Результирующее дерево решений

Представим третью ветвь дерева с помощью применения подстановочных операций и импликации:

$$x_1^m(x_3^c(P)) \to x_5^b;$$

$$x_2^1(x_1^f(x_3^c(P))) \to x_5^b;$$

$$x_2^0(x_1^f(x_3^c(P))) \to x_5^t.$$

Теперь можно ответить на поставленный вопрос. Например, женщина, выбравшая дешевую путевку, имея один автомобиль и средний уровень дохода, предположительно будет путешествовать на автобусе.

Заключение

Мы показали, что логические уравнения алгебры конечных предикатов могут быть использованы для определения связей между различными атрибутами при построении деревьев решений. Процедура определения искомых логических правил, связывающих целевую переменную и остальные атрибуты, сводится к навешиванию квантора существования на все переменные, не участвующие в итоговом правиле. При исключении переменных с помощью квантора исходная формула упрощается, что дает возможность получения простых и удобных для интерпретации зависимостей. В сочетании с известными методами генерации деревьев логико-алгебраический подход позволяет анализировать структуры данных в виде СДНФ и ДНФ, а также более сложные логические структуры.

Список литературы

1. Шабанов-Кушнаренко Ю.П. Теория интеллекта. Технические средства: монография / Ю.П. Шабанов-Кушнаренко. – Х. : Вища шк., 1986. – 134 с.

2. Шабанов-Кушнаренко Ю.П. Теория интеллекта. Математические средства: монография / Ю.П. Шабанов-Кушнаренко. – Х. : Вища шк., 1984. – 142 с.

3. Sitnikova P.E. Discovering Salient Data Features Based on Composing and Manipulating Logical Equations / P.E. Sitnikova, D.E. Sitnikov, B. D'Cruz // Data Mining II: Proc. 2nd International Conference.– Cambridge: WIT Press, 2000. – P. 241-248.

Поступила в редколлегию 23.03.2017

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

ЛОГІКО-АЛГЕБРАЇЧНИХ ПІДХІД ДО ПОБУДОВИ ДЕРЕВ РІШЕНЬ ТА ОПЕРАЦІЙ З НИМИ

Д.Е. Ситніков, П.Е. Ситнікова

У цій статті запропоновано логіко-алгебраїчний підхід до вилучення логічних правил з баз даних за допомогою побудови дерев рішень. Проаналізовані основні етапи генерації правила. Запропонований підхід дозволяє працювати з базами даних і базами знань у вигляді логічних рівнянь. Розглянуто операцію квантор існування в якості основної операції з вилучення нецільової змінної.

***Ключові слова**: дерева рішень, CART-алгоритм, ID3-алгоритм, інформаційна ентропія, індекс Gain.*

A LOGIC-ALGEBRAIC APPROACH TO BUILDING DECISION TREES AND OPERATIONS WITH THEM

D. Sitnikov, P. Sitnikova

In this paper, we suggest a logic-algebraic approach to the extraction of logic rules from databases with the help of building decision trees. The main stages of rule generation have been analyzed. The suggested approach allows working with relational databases and knowledge bases in the form of logic equations. The existence quantifier operation has been considered as the main operation for non-target variable elimination.

***Keywords**: Decision trees, CART-algorithm, ID3-algorithm, information entropy, Gain index.*