МЕТОДЫ И АЛГОРИТМЫ ГЕНЕРАЦИИ АССОЦИАТИВНЫХ ПРАВИЛ ПРИ АНАЛИЗЕ ДАННЫХ ПО ГАРАНТИЙНЫМ ОБЯЗАТЕЛЬСТВАМ АВТОТРАНСПОРТА

Нгуен Дык Тхань

Метод и алгоритм генерации ассоциативных правил основываются на теории элементарного множества и методах обработки базы данных для выявления полезных взаимосвязей между атрибутами продукции (автотранспорт) и причинами неисправностей (например, в формате кодов ремонтных работ). Как уже было показано в п. 3.1, каждое отношение представлено как ассоциативное правило, состоящее из 2-х множеств операторов: условие и решение. Правило в самой общей форме может быть представлено как: ЕСЛИ… <операторы условия>, ТО… <операторы решения>. Часть правила ЕСЛИ, включает множество атрибутов, представляющих условия – характеристики автотранспорта (например, даты изготовления и ремонта, пробег до возникновения гарантийного случая, типы трансмиссии и двигателя, и так далее), а часть правила ТО включает множество атрибутов, представляющих результаты решения (например, конкретные коды работ). После разработки ассоциативных правил, алгоритм использует методы статистического анализа для оценки значимости каждого правила. Правила, прошедшие проверку значимости, приводятся в решении.

Перед описанием шагов алгоритма генерации ассоциативного правила, введены следующие обозначения:

$C = \{c_1, c_2, \ldots, c_n\}$ – множество атрибутов условия;

$D = \{d_1, d_2, \ldots, d_m\}$ – множество атрибутов решения;

C_i – элементарное множество C, где $i = 1, ..., p$;

D_j – элементарное множество D, где $j = 1, ..., q$;

$V(C_i, c_k)$ – значение атрибута c_k в элементарном множестве C_i;

$V(D_j, d_l)$ – значение атрибута d_l в элементарном множестве D_j;

X_{ij} – пересечение элементарных множеств C_i и D_j;

$f(r, a_j)$ – значение атрибута a_j для объекта r;

P – процент объектов в элементарном множестве атрибутов условия, которые соответствуют правилу;

Q – процент объектов в элементарном множестве атрибутов решений, которые соответствуют правилу;

N – общее количество объектов во множестве данных;

RS – относительная сила (Relative Strength), процент от объектов, которые соответствуют правилу.

Блок-схема алгоритма генерации ассоциативных правил при анализе данных по гарантийным обязательствам автотранспорта показана на рисунке 1.

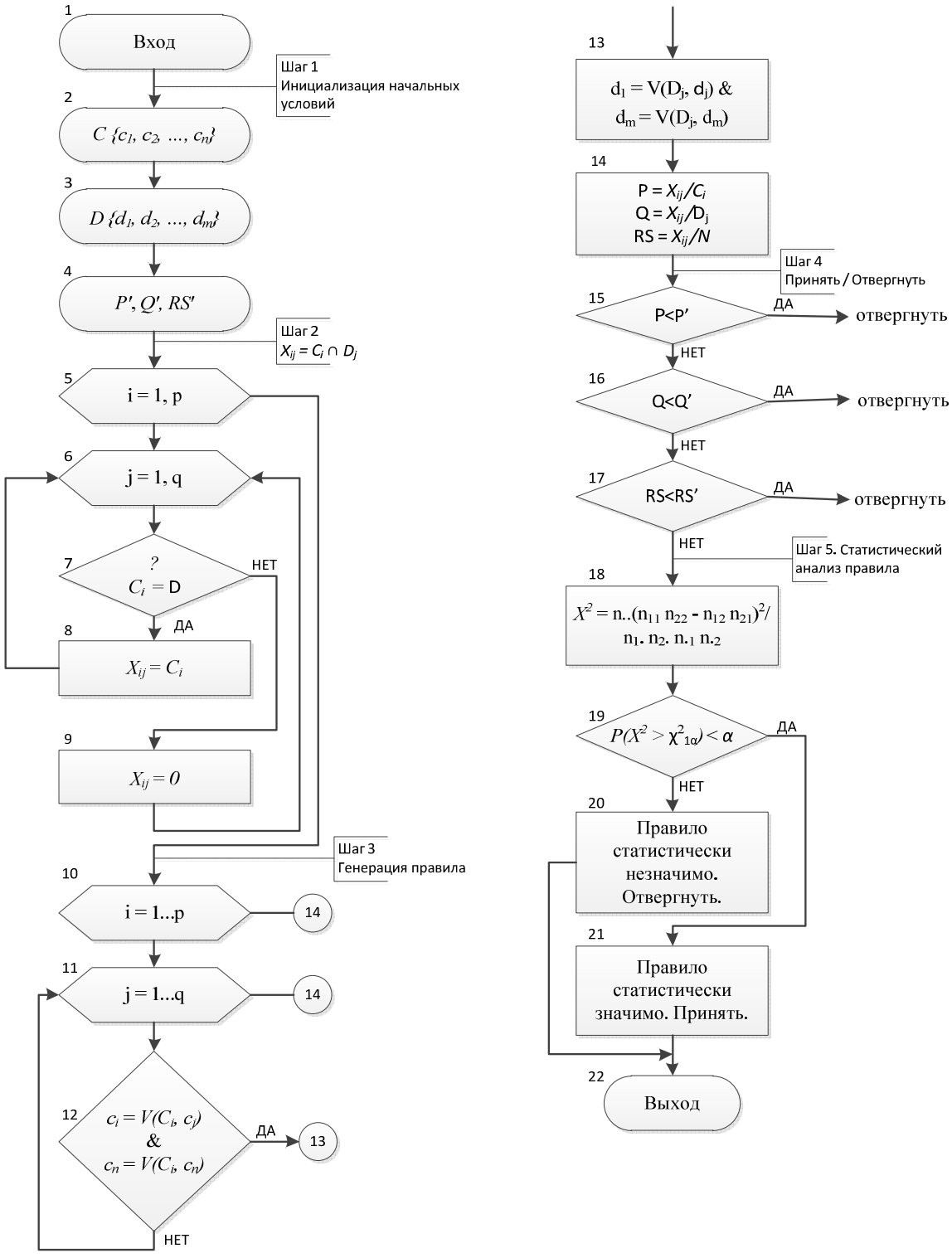

Рисунок 1. - Блок-схема алгоритма генерации ассоциативных правил при анализе данных по гарантийным обязательствам автотранспорта

Шаги алгоритма генерации ассоциативных правил

Шаг 1. Инициализировать $C = \{c_1, c_2, \ldots, c_n\}$; $D = \{d_1, d_2, \ldots, d_m\}$ и определить P', Q' и RS', как пороги для P, Q, и RS, соответственно.

Шаг 2. Определить $X_{ij} = C_i \cap D_j$ для каждого $i = 1, \ldots, p$ и $j = 1, \ldots, q$.

Шаг 3. Для каждого $X_{ij} \neq Ø$ генерировать правило:

ЕСЛИ $c_1 = V(C_i, c_1)$ И … И $c_n = V(C_i, c_n)$,

ТО $d_1 = V(D_j, d_1)$ И … И $d_m = V(D_j, d_m)$ [P, Q, RS],

где $P = |X_{ij}|/|C_i|$; $Q = |X_{ij}|/|D_j|$; $RS = |X_{ij}|/N$.

Шаг 4. Отвергнуть правила полученные на 3-м шаге, для которых $P < P'$ или $Q < Q'$ или $RS < RS'$.

Шаг 5. Для каждого из оставшихся правил, выполнять статистический анализ с использованием статистики хи-квадрат X^2.

Для каждого правила, полученного на 4-м шаге, необходимо сформировать таблицу 1 для оценки достоверности ассоциативных правил. Ассоциативное правило значимо, как показано в таблице 1.

Таблица 1 – Оценка достоверности ассоциативных правил

Условия правила (ЕСЛИ)	Решения правила (ТО)		
	$d_l = V(D_j, d_l)$, $\forall\ l = 1,\ldots, m$	По крайней мере одно $d_l \neq V(D_j, d_l)$	Всего
$c_k = V(C_i, c_k)$, $\forall k = 1,\ldots, n$	$n_{11} = \|X_{ij}\|$	$n_{12} = \|C_i\| - \|X_{ij}\|$	$n_{1\blacksquare} = \|C_i\|$
По крайней мере одно $c_k \neq V(C_i, c_k)$	$n_{21} = \|D_j\| - \|X_{ij}\|$	$n_{22} = N - \|C_i\| - \|D_j\| + \|X_{ij}\|$	$n_{2\blacksquare} = N - \|C_i\|$
Всего	$n_{\blacksquare 1} = \|D_j\|$	$n_{\blacksquare 2} = N - \|D_j\|$	$n_{\blacksquare\blacksquare} = N$

Используя информацию из таблицы 1, для правила рассчитывается статистика хи-квадрат с одной степенью свободы:

$X^2 = n..(n_{11}\, n_{22} - n_{12}\, n_{21})^2/n_1.\, n_2.\, n._1\, n._2$.

ЕСЛИ р-значения $P(X^2 > \chi^2_{1\alpha}) < \alpha$, где α является уровень значимости теста, ТО правило можно принять (т.е. правило является статистически значимым на уровне достоверности α). В противном случае надо перейти к шагу 5.

Алгоритм генерации ассоциативного правила начинается с инициализации множеств C и D и использования выбранных пороговых значений P', Q' и RS'. Каждое непустое пересечение элементарных множеств C и D, полученных на 2-м шаге, представляет собой единственное ЕСЛИ-ТО решение правила на 3-м шаге. В этом правиле решения часть правила ЕСЛИ включает в себя множество атрибутов, представляющих условия правила и часть правила ТО, включает множество атрибутов, представляющих решения. Для каждого правила, параметры P, Q, и RS являются также оценками на этом шаге. Правила полученные на шаге 3 далее проходят проверку на шаге 4 с использованием выбранных пороговых значений параметров правила P', Q' и RS'. Наконец, на 5-м шаге, правила, которые удовлетворяют условиям шага 4 оцениваются с помощью методов статистического анализа. Для того, чтобы проиллюстрировать шаги алгоритма правила решения и того, как параметры P, Q и RS используются для анализа правила, рассмотрим пример гарантийных данных, приведенных в таблице 2.

Таблица 2 – Образец множества гарантийных данных для иллюстративного примера

№ объекта	Двигатель	Трансмиссия	Год выпуска	Код работы
1	Б	2	2007	3
2	Б	2	2008	1
3	А	2	2008	3
4	А	2	2007	3
5	Б	1	2007	2
6	А	1	2009	2
7	А	1	2009	3
8	А	2	2009	3
9	Б	1	2007	2
10	Б	2	2009	2
11	А	1	2008	3
12	Б	2	2008	1

Предположим, что в анализе данных применяется алгоритм генерации ассоциативного правило для выявления возможных связей между атрибутами продукции "Двигатель" и "Код работы", на основе множества данных в таблице 3.2. На 1-м шаге алгоритма, множества C = {двигатель} и D = {код работы} инициализированы и определены следующие уровни пороговых значений P' = 50%, Q' = 50%, и RS' = 25%. На 2-м шаге, определяются элементарные множества C и D и рассчитываются их соответствующие пересечения:

C_1 = {1, 2, 5, 9, 10, 12}, C_2 = {3, 4, 6, 7, 8, 11};

D_1 = {1, 3, 4, 7, 8, 11}, D_2 = {2, 12}, $D3$ = {5, 6, 9, 10};

$X_{11} = C_1 \cap D_1$ = {1}, $X_{12} = C_1 \cap D_2$ = {2, 12}, $X_{13} = C_1 \cap D_3$ = {5, 9, 10};

$X_{21} = C_2 \cap D_1$ = {3, 4, 7, 8, 11}, $X_{22} = C_2 \cap D_3 = \emptyset$, $X_{23} = C_2 \cap D_3$ = {6}.

Помимо этого, на этом шаге определяются значения элементарных множеств:

$V(C_1$, “Двигатель”) = B, $V(C_2$, “Двигатель”) = A;

$V(D_1$, “Код работы”) = 3, $V(D_2$, “Код работы”) = 1 и $V(D_3$, “Код работы”) = 2.

На 3-м шаге алгоритма, формируются правила принятия решений и для каждого правила рассчитываются значения параметров P, Q, и RS:

Правило 1: ЕСЛИ двигатель = B, ТО код работ = 3. [P = 16,67%, Q = 16,67%, RS = 8,33%].

Правило 2: ЕСЛИ двигатель = B, ТО код работ = 1. [P = 33,33%, Q = 100%, RS = 16,67%].

Правило 3: ЕСЛИ двигатель = B, ТО код работ = 2. [P = 50%, Q = 75%, PC = 25%].

Правило 4: ЕСЛИ двигатель = A, ТО код работ = 3. [P = 83,33%, Q = 83,33%, RS = 41,67%].

Правило 5: ЕСЛИ двигатель = A, ТО код работ = 2. [P = 16,67%, Q = 25%, RS = 8,33%].

В полученной группе правил, Правило 1 соответствует объектам 1 и 4 в таблице 3.2, а Правилу 2 соответствуют объекты 2 и 12. Значение P = 33,33% для Правила 1 означает, что данное правило распространяется на одну треть объектов из множества данных, соответствующих условию Двигатель = *B*. Значение Q = 100% в том же правиле означает, что это правило охватывает все объекты с решением Код работ = 1. RS = 16,67% означает, что только 2 из 12 объектов соответствуют этому правилу.

На 4-м шаге, Правила 1, 2 и 5 отклоняются, поскольку, по крайней мере, один из параметров поддержки правил имеет значение ниже порогового, в то время как правила 3 и 4 сохраняются в качестве потенциальных строгих правил. Наконец, на шаге 5, тест хи-квадрат показывает, что при α = 0,05, Правило 4 является существенным с X^2= 5,33 и p-значением = 0,02, в то время как правило 3 не имеет существенного значения, с X^2= 1,5 и p-значения = 0,22. В таблице 3.3 приведены пояснения того, как на 5-м шаге алгоритма данные в таблице 3.2 табулированы в таблицу 3 для правила 3 и 4.

Необходимо отметить, что для ассоциативного правила, включающего в себя, по крайней мере, два атрибута решения, из которых только один в значительной степени связан с атрибутом (атрибутами) условий, тест хи-квадрат может показать, что это правило является существенным. Это схоже с проблемой, когда хи-квадрат используется для поиска связи между двумя факторами на более чем двух уровнях. Невысокое p-значение указывает на то, что два фактора связаны без выявления предоставления знаний о том, какой уровень (уровни) какого фактора описывает связь между ними.

Таблица 3 – Генерирование правил 3 и 4

Правило 3	Код работы=2	Код работы≠2	Всего
Двигатель=Б	3	3	6
Двигатель≠Б	1	5	6
Всего	4	8	12
Правило 4			
Двигатель=А	5	1	6
Двигатель≠А	1	5	6
Всего	6	6	12

В алгоритме генерации ассоциативного правила оценки элементарных множеств и их пересечений требуют значительного времени вычислений. Алгоритм предполагает, что таблицы данных хранятся в плоском файле и обрабатываются строка за строкой. Такой подход может потребовать значительного времени вычислений при использовании для анализа большого множества данных. Однако, в большинстве DM-приложений гарантийные данные хранятся в среде базы данных, в которой можно использовать операции эффективной множество-ориентированной базы данных, такие как projection (проекция данных) и cardinality (мощность множества) [127], для выполнения вычислений. Например, для таблицы с атрибутами $A=\{A_1, A_2, \ldots, A_n\}$ операции *проекции* в атрибуты $A' \subseteq A$ позволяет быстро генерировать все возможные

комбинации значений атрибутов A′ и оценки числа строк в исходной таблице, которые имеют эти значения атрибутов. Операция *проекция* может эффективно использоваться для получения всех возможных непустых пересечений элементарных множеств в алгоритме генерации ассоциативного правила. Для иллюстрации его применения, рассмотрим данные в таблице 3.2 и результаты операции проецирования для атрибутов "Двигатель" и "Код работы", приведенные в таблице 4.

Таблица 4 - Результаты операции проецирования

Двигатель	Код работы	Мощность
А	2	1
А	3	5
Б	1	2
Б	2	3
Б	3	1

Можно установить, что комбинации {*A*, 2}, {*A*, 3}, {*B*, 1}, {*B*, 2} и {*B*, 3} представляют уникальное значение комбинации для атрибутов "Двигатель" и "Код работ" в таблице 2. В графе "Мощность" в таблице 4 показано количество строк в таблице 2, которые имеют заданную комбинацию значений атрибутов. Каждому уникальному сочетанию значений атрибутов соответствует непустое пересечение элементарных множеств атрибутов "Двигатель" и "Код работ", и одно из пяти полученных ранее правил. Следовательно, все пять правил могут быть получены из ряда проекции таблицы 3.4. Например, можно видеть, что правила 1 и 2 могут быть получены с использованием пятого и третьего ряда таблицы проецирования, соответственно.

Таким образом, в алгоритме генерации ассоциативного правила, вместо вычисления элементарных множеств $C \cup D$ и использования пересечений множества для создания правил, можно получить те же правила с проекцией исходной таблицы в атрибуты $C \cup D$. Так как этот подход использует операции высокоэффективной множественно-ориентированной базы данных, процесс генерации правила обеспечивает высокую скорость и масштабируемость для больших множеств данных. Следовательно, шаги 2 и 3 оригинального алгоритма могут быть модифицированы, чтобы воспользоваться операциями проецирования, следующим образом:

Шаг 2. Создание проекции исходной таблицы в атрибуты $C \cup D$.

Шаг 3. Для каждой строки r в результирующую таблицу, генерировать правило вида:

ЕСЛИ $c_1 = f(r, c_1)$ И … И $c_n = f(r, c_n)$;

ТО $d_1 = f(r, d_1)$ И … И $d_m = f(r, d_m)$ [P, Q, RS];

где – $P = |C_r \cap D_r|/|C_r|$; $Q = |C_r \cap D_r|/|D_r|$; $RS = |C_r \cap D_r|/N$; $S = |C_r \cap D_r|$;

C_r и D_r – элементарные множества C и D, соответственно;

$|C_r \cap D_r|$, $|C_r|$, и $|D_r|$ – получаются из проекции и мощности данных таблицы атрибутов $C \cup D$, C и D соответственно.

Таким образом, получены DM-метод и алгоритм генерации ассоциативных правил при анализе данных по гарантийным обязательствам автотранспорта. Алгоритм использует теорию элементарного множества и методы обработки базы данных для раскрытия необходимых отношений между атрибутами автомобиля и причинами неисправности. Эти отношения (знания) представлены с использованием логических конструкций ЕСЛИ-ТО ассоциативных правил, где ЕСЛИ – часть правила, включающая множество атрибутов, представляющих характеристики автотранспорта (например, даты изготовления и ремонта, пробег на момент ремонта, тип трансмиссии и двигателя и так далее), а часть правила ТО включает множество атрибутов, которые представляют результаты решения (например, задачи связанные с кодом работ).

На основе полученного алгоритма генерации ассоциативных правил разработан DM-алгоритм поиска последовательных шаблонов в потоках данных по гарантийным обязательствам в автомобильной промышленности.

Список информационных источников

[1] Нгуен Д.Т. Автоматизация анализа данных для принятия решении в режиме реального времени в автомобильной промышленности // Вестник МАДИ. – 2012. – № 1. – с. 104-109.

[2] Нгуен Д.Т. Автоматизации поддержки решений по управлению гарантийными обязательствами в автомобильной промышленности // Вестник МАДИ. – 2012. – № 3. – с. 74-80.

[3] Нгуен Д.Т. Метод и алгоритм генерации ассоциативных правил при анализе данных по гарантийным обязательствам автотранспорта. // В мире научных открытий. – 2012. – №2.6(26). – с. 138-144.

[4] Нгуен Д.Т. Метод и алгоритм поиска последовательных шаблонов в гарантийных данных автомобильной промышленности // В мире научных открытий. – 2012. – №2.6(26). – с. 45-52.

[5] Нгуен Д.Т. Применение ассоциативных правил Data Mining в анализе гарантийных данных по гарантийным обязательствам автотранспорта // Грузовик. – 2012. – №3. – с. 36-44.