



Ассоциативные правила



Описание

Компонент выявляет ассоциативные правила в транзакционных данных. Примером такого правила служит утверждение, что покупатель, приобретающий 'Хлеб' (*условие правила*), купит и 'Молоко' (*следствие правила*) с вероятностью 75%. Транзакцией в данном примере является чек продажи, содержащий список приобретенных товаров, а каждый товар в чеке является элементом транзакции. При поиске ассоциативных правил применяется алгоритм FP-Growth.

Наряду с анализом основных данных транзакций возможно учитывать и вспомогательные. Например, если транзакцией является чек, а элементами — товары, то в качестве вспомогательных данных могут быть использованы: пол покупателя, возраст, регион, сезон и т.д. Фактически вспомогательные данные рассматриваются алгоритмом как еще одни элементы транзакций, и обозначение "вспомогательные" они имеют лишь в контексте аналитической задачи. Поскольку вспомогательные данные чаще представляются в источниках данных как дополнительные атрибуты транзакций, узел имеет отдельный вход для их приема.

Для получения результирующих наборов требуется предварительное обучение узла.

Вход

-  — **Входной источник данных** (таблица данных). Необязательный.
Порт ожидает набор данных со столбцами идентификаторов и элементов транзакций.
-  — **Вспомогательные данные** (таблица данных). Необязательный.
Принимает дополнительные элементы транзакций, которые возможно учитывать при расчете ассоциативных правил.

Требования к принимаемым данным

▼ Примеры входных наборов:

Входной источник данных:

Идентификатор транзакции	Элементы транзакции
Чек №000001	Хлеб
Чек №000001	Молоко
Чек №000001	Масло

Идентификатор транзакции	Элементы транзакции
Чек №000002	Хлеб
Чек №000003	Хлеб
Чек №000003	Молоко

В отличие от таблицы, принимаемой портом "Входной источник данных", элементы транзакций вспомогательных данных должны располагаться в строках, а не в столбцах. Таким образом, структура таблицы предполагает наличие одного поля идентификатора транзакций и одного или более полей элементов транзакций.

Вспомогательные данные:

Идентификатор транзакции	Пол покупателя	Возраст	Регион	Сезон
Чек №000001	м	20-30	Московская обл.	1
Чек №000002	ж	40-50	Калининградская обл.	2
Чек №000003	ж	30-40	Орловская обл.	1

В настройках портов следует выставить параметр "Назначение" для полей участвующих в обработке. Параметр может принимать значения: *Неиспользуемое*, *Транзакция*, *Элемент*. Идентификаторы и элементы транзакций могут быть представлены только дискретными данными.

Выход

- **Популярные наборы** (таблица данных). Наборы элементов, наиболее часто встречающиеся в транзакциях (частые наборы).
- **Ассоциативные правила** (таблица данных). Выявленные ассоциативные правила и их показатели: поддержка, достоверность, лифт.
- **Применение правил** (таблица данных). Содержит транзакции входного набора данных, в которых срабатывают выявленные правила.

Мастер настройки

Включает следующие группы параметров:

Частые наборы

Задаются условия, по которым определяются *частые предметные наборы* — наборы элементов, наиболее часто встречающиеся в транзакциях. В дальнейшем только эти наборы участвуют в формировании правил:

- **Минимальная поддержка, %** — минимальная частота, с которой набор встречается в транзакциях (значение 0 до 100).
- **Исключать элементы с поддержкой, больше максимальной** — элементы, которые слишком часто встречаются в транзакциях, как правило, не несут информации о закономерностях сочетания с ними других элементов. Для их определения и исключения из частых наборов задается:
 - **Максимальная поддержка, %** — максимальная частота, с которой элемент встречается в транзакциях (значение от 0 до 100).
- **Содержащие выбранные элементы** — задает поля вспомогательного набора данных, содержащие дополнительные элементы транзакций.
- **Исключать одиночные наборы** — исключает наборы из одного элемента;
- **Максимальное число элементов** — задает максимальное количество элементов в наборе (максимальная мощность набора).

Ассоциативные правила

В результирующий набор попадают правила, удовлетворяющие следующим условиям:

- **Минимальная достоверность правила, %** — позволяет отсеять наименее точные правила (значение от 0 до 100).
- **Минимальный лифт правила** — значение лифта > 1 косвенно подтверждает значимость правила, поскольку говорит о положительной связи двух предметных наборов (условия и следствия правила). Значение лифта, равное или меньшее 1, говорит об отсутствии или отрицательной связи. Задавая минимальную величину лифта, можно отсеять наименее значимые правила.
- **Максимальное число следствий** — максимальное количество элементов в наборе, представляющем следствие правила.