

## Кластеризация транзакций

### Описание

В основе кластеризации транзакций лежит алгоритм CLOPE, применение которого позволяет обрабатывать огромные массивы транзакционных данных: чеки в супермаркетах, логи посещений веб-ресурсов и другие. Задача состоит в получении такого разбиения всего множества транзакций, чтобы похожие транзакции оказались в одном кластере, а отличающиеся друг от друга — в разных кластерах.

Алгоритм автоматически подбирает количество кластеров. Аналитик может повлиять на результат с помощью коэффициента отталкивания и назначения максимального числа кластеров или итераций.

Для получения результирующих наборов требуется предварительное обучение узла.

### Порты

#### Вход

-  — Входной источник данных (таблица данных). Необязательный.
-  — Дополнительный вход (таблица данных). Необязательный.

Входные данные - это транзакции, сформированные в 2 поля, одно из которых является транзакцией, а второе - элементом. Например, первое поле - код покупателя, второе - список его покупок.

#### Требования к принимаемым данным

Поля с непрерывным видом данных непригодны для обработки.

Преобразовать к требуемому виду (с полями Транзакция и Элемент) обычный набор данных, состоящий из столбцов с измеренными свойствами объектов всегда можно при помощи узлов Loginom, таких как Свёртка столбцов и Калькулятор

#### Выход

-  — **Разбиение на кластеры.**

Таблица данных, состоящая из:

- Поля, назначение которого — «Транзакция».

- Номеров кластеров — каждой транзакции присвоен номер того кластера, в который она входит.
-  — **Параметры кластеров.**

Таблица данных, состоящая из:

- Номеров кластеров — перечислены номера сформированных кластеров.
- $N$  — количество транзакций, вошедших в кластер.
- $W$  — ширина кластера, т.е. число уникальных объектов в нем.
- $S$  — мощность (площадь) кластера.

## Мастер настройки

На данном этапе необходимо задать:

- **Поля для кластеризации.**
  - Для полей, участвующих в обработке, выставить назначения «Транзакция» и «Элемент».
  - Для прочих полей оставить «Не задано».
- **Коэффициент отталкивания** — с его помощью регулируется уровень сходства транзакций внутри кластера, и, как следствие, финальное количество кластеров. Чем больше коэффициент, тем ниже уровень сходства и тем больше кластеров будет сгенерировано. По умолчанию значение коэффициента отталкивания установлено 2,6. Диапазон изменения значений от 1 до 4.
- **Ограничить число кластеров** — предполагает возможность вручную задать наибольшее количество кластеров, которые может выдать в результате алгоритм. Этой настройкой можно воспользоваться в случае, если задача требует определенного числа кластеров, например, не больше 15.
- **Ограничить число итераций** — искусственное прекращение работы алгоритма в случае, если количество итераций алгоритма превышает заданное максимальное число итераций. В алгоритме первый проход по таблице транзакций служит для построения начального разбиения, определяемого функцией стоимости, после чего для повышения качества кластеризации и оптимизации функции дополнительно сканируются таблицы несколько раз, пока изменения в разбиении не прекратятся. Ограничивать максимальное количество итераций следует в случае большого количества данных для предварительной оценки качества кластеризации.