

 Кластеризация

Описание

Кластеризация (сегментация) — это группировка объектов (наблюдений, событий) на основе данных, описывающих свойства объектов. Объекты внутри кластера должны быть похожими друг на друга и отличаться от других, которые вошли в другие кластеры.

Компонент производит кластеризацию объектов на основе алгоритмов k-means и g-means. Основное отличие одного алгоритма от другого в том, известно ли заранее количество кластеров. Если количество кластеров известно, то применяется алгоритм *k-means*, в противном случае — *g-means*, который определяет это количество автоматически в рамках заданного интервала.

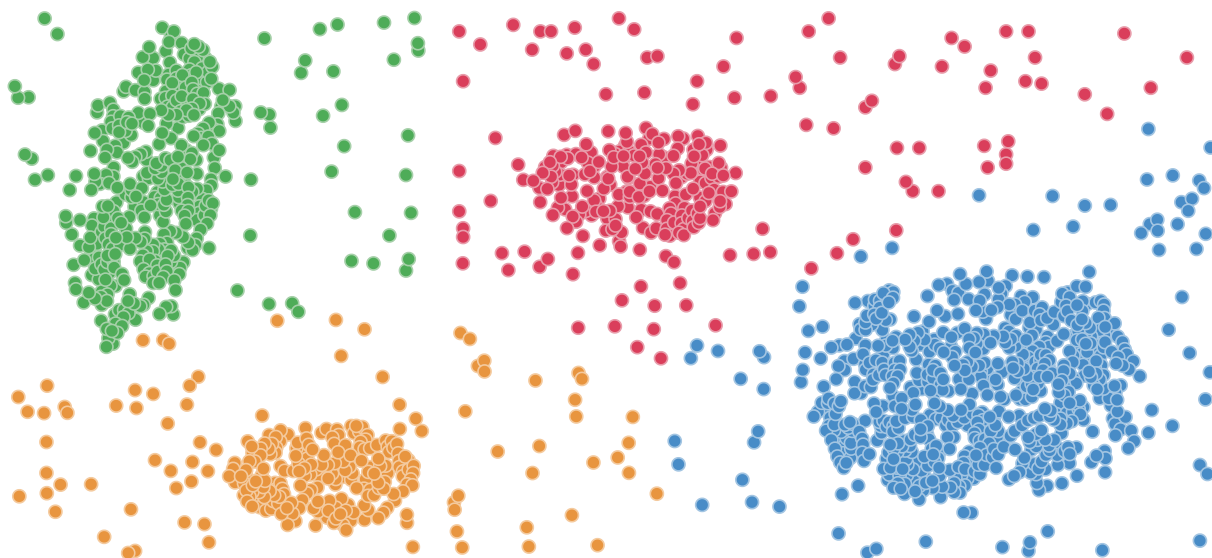


Рисунок 1. Иллюстрация работы алгоритма *k-means*

На рисунке цветом выделены отдельные кластеры и объекты, им принадлежащие.

Для получения результирующих наборов требуется предварительное обучение узла.

Порты

Вход

-  Входной источник данных (таблица данных).

Требования к принимаемым данным

Поле будет запрещено к использованию, если:

- оно является дискретным и содержит всего одно уникальное значение;
- оно непрерывное и с нулевой дисперсией;
- оно содержит пропущенные значения.

Выход

-  Разбиение на кластеры (таблица данных).

Таблица, состоящая из полей:

- **Номер кластера** — каждому объекту присвоен номер того кластера, в который он входит.
- **Расстояние до центра кластера** — положение объекта относительно центра кластера.
- Поля исходного набора данных (значения не изменяются).
-  Центры кластеров (таблица данных).

Центр кластера — среднее значение переменных объектов, входящих в кластер. Результат — таблица, количество записей которой соответствует числу кластеров, т.е. данные сгруппированы по кластерам. Состоит из полей:

- **Номер кластера** — перечислены номера сформированных кластеров.
- Поля исходного набора данных, в ячейках которых рассчитано среднее значение параметров.

Мастер настройки узла

Мастер настройки включает в себя следующие группы параметров:

- Настройка входных столбцов;
- Настройки нормализации;
- Кластеризация.

Настройка входных столбцов

- Выбор полей для кластеризации:

- Для полей, участвующих в кластеризации, выставить назначение *Используемое*.
- Для прочих полей оставить *Не задано*.

Кластеризация

- При заданном числе кластеров:
 - Снять галочку *Автоопределение числа кластеров*.
 - Ввести нужное количество кластеров (должно быть больше 2). По умолчанию — 3.
- При автоматическом определении числа кластеров:
 - Минимальное число кластеров. По умолчанию — 1.
 - Максимальное число кластеров. По умолчанию — 10.
 - Порог разделения кластеров (в интервале от 0,1 до 5). Чем больше порог разделения, тем больше кластеров будет сгенерировано при кластеризации.

Random seed — начальное число (целое, положительное), которое используется для инициализации генератора псевдослучайных чисел. Последовательность чисел генератора полностью определяется начальным числом. Если генератор повторно инициализируется с тем же начальным числом, он выдаст ту же последовательность чисел.

Параметр влияет на воспроизводимость результата обучения. Можно повторить результат обучения узла, если подать те же данные и выставить тот же random seed.

Для параметра доступны следующие команды:

- Всегда случайно — начальное число всегда будет случайным.
- Генерировать — сгенерируется новое начальное число.
- Копировать — в буфер обмена будет скопировано указанное значение.