



Разбиение на множества

Описание

Разбиение на множества применяется в случае, когда задача анализа требует разделить исходный набор данных на обучающую и тестовую выборки. Размер этих выборок можно настроить, а отбор записей для них производится на основе выбранного метода сэмплинга. Сначала отбираются записи для обучающей выборки, оставшиеся используются для тестовой (этот порядок можно изменить в мастере настройки узла).

Порты

Вход

-  Входной источник данных (таблица данных).

Выход

-  Общий выходной набор (таблица данных). Содержит все строки, которые были взяты для обеих выборок. Добавляется поле "тестовое множество": в этом поле значение "true" свидетельствует о том, что запись попала в тестовую выборку, а значение "false" — в обучающую.
-  Обучающий выходной набор (таблица данных).
-  Тестовый выходной набор (таблица данных).

Мастер настройки

- **Состояние входа** — при активном состоянии позволяет использовать входные данные. Например, в "Отбор со смещением" нужны данные из входного набора.
- **Общее число записей** — число записей таблицы входного источника данных.
- **Область настройки количества строк для обучающей и тестовой выборок.**
Размер каждой выборки задается вручную; счетная кнопка "способ" предоставляет выбор — непосредственно указать количество строк или задать его в процентах от размера исходной таблицы. Сумма строк выборок не может быть больше, чем количество строк в исходной таблице. Если обучающее и тестовое множество не согласуются друг с другом по количеству записей (в сумме дают количество строк больше строк входного набора), то первым формируется

множество согласно флагу *Приоритет тестового множества*, второе множество формируется по остаточному принципу.

- **Метод сэмплинга:**

- *Случайный* — записи случайным образом выбираются из исходного набора данных и помещаются в результирующую выборку.
- *Равномерный случайный* — все записи исходного набора данных разделяются на группы и затем из каждой группы случайным образом выбираются и помещаются в результирующую выборку. Размерность группы задается в параметрах метода.
- *Стратифицированный* — все записи исходного набора данных разделяются на однородные группы (страты), после чего из каждой группы случайным образом выбираются и помещаются в результирующую выборку. Поля, определяющие страты, задаются в параметрах метода с помощью флажков.
- *Последовательный* — из исходного набора данных последовательно выбираются записи и помещаются в результирующую выборку. Размеры сэмплированного и неиспользованного множеств настраиваются в параметрах метода.
- *Отбор со смещением* — в исходном наборе данных сокращается или увеличивается количество записей с выбранными уникальными значениями перед обработкой. В параметрах этого метода коэффициент увеличения задается в поле "фактор" напротив каждого уникального значения выбранного столбца исходной таблицы, также число записей для каждого уникального значения можно ввести вручную.

- **Приоритет тестового множества (необязательный флаг).**

Активация данного флага приводит к тому, что сначала отбираются записи для тестовой выборки, а оставшиеся используются для обучающей.

Имеется три режима отбора, определяемые параметром "Положение приоритетного тестового множества":

- *Определяется алгоритмом* — записи будут отобраны в соответствии с выбранным ранее методом сэмплинга.
- *В начале набора* — в качестве тестовой выборки будут использованы строки из начала набора, взятые в том же порядке, что и в исходной таблице.
- *В конце набора* — в качестве тестовой выборки будут использованы строки из конца набора, взятые в том же порядке, что и в исходной таблице.

- **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 порядок случайного разбиения на тестовое и обучающее множество. Можно повторить разбиение, 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.