



Сэмплинг

Описание

Компонент реализует различные способы отбора записей в выборку из исходного набора данных. Такая выборка создается для сокращения анализируемого объема данных, например, во избежание избыточных вычислений, т. к. трудоемкость некоторых алгоритмов Data Mining экспоненциально возрастает с увеличением объема данных.

При формировании выборки пользователь может задать метод отбора: случайный, равномерный случайный, стратифицированный, последовательный, отбор со смещением.

Порты

Вход

-  Входной источник данных (таблица данных).

Выход

-  Выходной набор данных (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 выходной таблицы.** Размер выходной таблицы задается вручную; кнопка-переключатель предоставляет выбор — непосредственно указать количество строк  или задать его в процентах  от размера исходной таблицы.
- **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. При установке флажка *Полнота списка уникальных значений* в результирующую выборку должны попасть записи из всех групп. Если обеспечить полноту списка окажется невозможным — возникнет ошибка.
- *Последовательный* — в параметрах метода определяется, откуда будет взята выборка — из начала исходной таблицы или из конца. Последовательность строк в выборке будет такой же, как и в исходной таблице.
- *Отбор со смещением* — при настройке параметров метода следует выбрать поле и уникальное значение из этого поля, для которого устанавливается фактор. Фактор, имеющий значение N, увеличит количество строк с интересующим нас уникальным значением в N раз в выходном наборе данных. N может принимать любые положительные значения. Следует учитывать, что максимальное количество уникальных значений в поле не должно превышать десять тысяч.
- **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 методы сэмплинга *Случайный* и *Равномерный случайный*. Повторить результат сэмплинга можно, если подать те же данные и указать 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.