



Конечные классы

Описание

Компонент Конечные классы предназначен для решения следующих задач:

- Преобразование непрерывных и дискретных входных полей, используемых для построения моделей бинарной классификации, путем квантования на основе метода совокупности доказательств (WoE-анализ). В результате каждое исходное значение признака заменяется на метку интервала квантования, в который данное значение попало. Использование результатов такого преобразования для построения моделей бинарной классификации (например, логистической регрессии) позволяет повысить их точность и устойчивость к изменению входных данных.
- Сокращение размерности данных за счет исключения признаков с низкой значимостью, снижения разнообразия значений признаков.
- Восстановление пропусков, когда пропуски образуют отдельную метку интервала квантования или объединяются с соседним, близким по значению WoE-индекса.
- Борьба с выбросами и экстремальными значениями — формирование меток интервала квантования при дискретизации непрерывного поля или объединение редких уникальных значений в одну категорию позволяет решить проблему экстремальных значений и выбросов.
- Упрощение описания исследуемых объектов.

Результатом работы узла *Конечные классы* является преобразование входных столбцов в последовательность интервалов, называемых *конечными классами*, каждому из которых присваивается определенная метка. Кроме этого, для каждого входного столбца может быть вычислен уровень значимости (отсутствует, очень низкая, низкая, средняя, высокая и очень высокая), на основе которого может производиться отбор переменных в модели бинарной классификации.

Вход

-  Входной источник данных (таблица данных).
-  Добавить еще один порт. Внешние диапазоны квантования (таблица данных).

Выход

-  Выходной набор данных (таблица данных).

▼ Структура данных:

I

- **?? Поля исходного набора данных** (значения не изменяются).
- **12 <Метка столбца> Номер класса** – идентификатор конечного класса, целое число (начиная с 0) – столбец создается всегда.
- **ab <Метка столбца> Метка** – метка конечного класса, полученная автоматическим путем (числовые границы, если это непрерывная переменная, или перечисление уникальных значений через «;», если переменная дискретная).
- **9.0 <Метка столбца> Значимость.**

-  Параметры классов (таблица данных).

▼ Структура данных:

- **12 Группа** – номер группы, к которой относится запись в таблице. Каждая группа записей ассоциирована с признаком (полем) исходного набора данных, являющимся входным для узла *Конечные классы*. Количество записей в группе соответствует числу конечных классов исходного столбца.
- **ab Идентификатор** – имя столбца, под которым он будет обрабатываться в наборе данных. Число столбцов равно числу входных полей узла *Конечные классы*.
- **ab Метка столбца** – мнемоническое обозначение входного столбца, под которым он будет виден пользователю в базе или хранилище данных. По умолчанию устанавливается название, под которым данный столбец виден в исходном наборе данных.
- **12 Номер класса** – порядковый номер (идентификатор), присвоенный классу при его формировании в узле *Конечные классы*.
- **ab Уник.значение** – для дискретных полей отображает их уникальные значения.
- **ab Метка класса** – метка класса, присвоенная ему при формировании в узле *Конечные классы*. Для числовых столбцов метка класса состоит из верхней и нижней границ класса (для нулевого класса указывается только нижняя граница с предлогом «от...», для класса с максимальным номером указывается верхняя граница с предлогом «до...»). Для категориальных полей, если каждый класс формируется для отдельной категории, то в качестве метки указывается эта категория. Если класс включает несколько категорий, то в метке перечисляются все категории, вошедшие в класс.
- **12 Число событий** – количество наблюдений в классе, для которых выходное значение – событие.
- **12 Число не-событий** – количество наблюдений в классе, для которых выходное значение – не-событие.
- **9.0 Доля событий** – отношение *Числа событий* к общему количеству *Числа событий* и *Числа не-событий*.
- **9.0 Доля не-событий** – отношение *Числа не-событий* к общему количеству *Числа событий* и *Числа не-событий*.
- **∞ Нижняя граница** – для числовых признаков указывается нижняя граница интервала числом. Для категориальных признаков нижняя граница обозначается двумя категориями – верхней категорией предыдущего класса и нижней категорией текущего.
- **∞ Верхняя граница** – для числовых признаков указывается верхняя граница интервала числом. Для категориальных признаков верхняя граница обозначается двумя категориями –

нижней категорией следующего класса и верхней категорией текущего.

- **9.0 Вес доказательства** – коэффициент *WoE* для каждого класса.
- **9.0 Информационный индекс** – указываются значения информационных индексов IV, вычисленные по каждому входному столбцу. Сумма частных информационных индексов по каждому классу дает общий информационный индекс признака, по которому определяется его значимость.
- **9.0 Доля класса** – отношение количества наблюдений в классе к общему числу наблюдений.
- **0/1 Верхняя граница диапазонов открыта.**
- **0/1 Предквантование** – показывает применялось ли предквантование в процессе формирования конечных классов.

-  Значимости столбцов (таблица данных).

▼ Структура данных:

- **ab Имя столбца** – идентификатор столбца, под которым он будет обрабатываться в наборе данных. Число столбцов равно числу входных полей узла *Конечные классы*.
- **ab Метка столбца** – мнемоническое обозначение входного столбца, под которым он будет виден пользователю в базе или хранилище данных. По умолчанию устанавливается название, под которым данный столбец виден в исходном наборе данных.
- **12 Число событий** – количество событий, попавших в данный класс.
- **12 Число не-событий** – количество не-событий, попавших в данный класс.
- **9.0 Доля событий** – отношение *Числа событий* к общему количеству *Числа событий* и *Числа не-событий*.
- **9.0 Доля не-событий** – отношение *Числа не-событий* к общему количеству *Числа событий* и *Числа не-событий*.
- **12 Всего** – общее число наблюдений в классе.
- **9.0 Информационный индекс** – указываются значения информационных индексов IV, вычисленные по каждому входному столбцу.
- **ab Значимость столбца** – уровень значимости входного столбца, определенный на основе *Информационного индекса*. Может принимать значения: отсутствует, очень низкая, низкая, средняя, высокая и очень высокая.

Мастер настройки

Мастер настройки состоит из следующих шагов:

- Настройка внешнего разбиения — появляется, если задан порт *Внешние диапазоны квантования*. Позволяет настроить параметры заранее настроенного внешнего разбиения.
- Настройка назначений столбцов — позволяет задать назначение столбцам, настройки входного и выходного полей, внешнего разбиения, а также настройки алгоритма для формирования конечных классов входных полей.

- Настройка конечных классов — позволяет просматривать начальные классы и результаты формирования конечных классов. Предназначена для внесения ручной корректировки в границы (или множества) сформированных конечных классов с целью достижения лучших результатов.
-

Статьи в разделе:

- Настройка внешнего разбиения
- Настройка назначений столбцов
- Настройка конечных классов