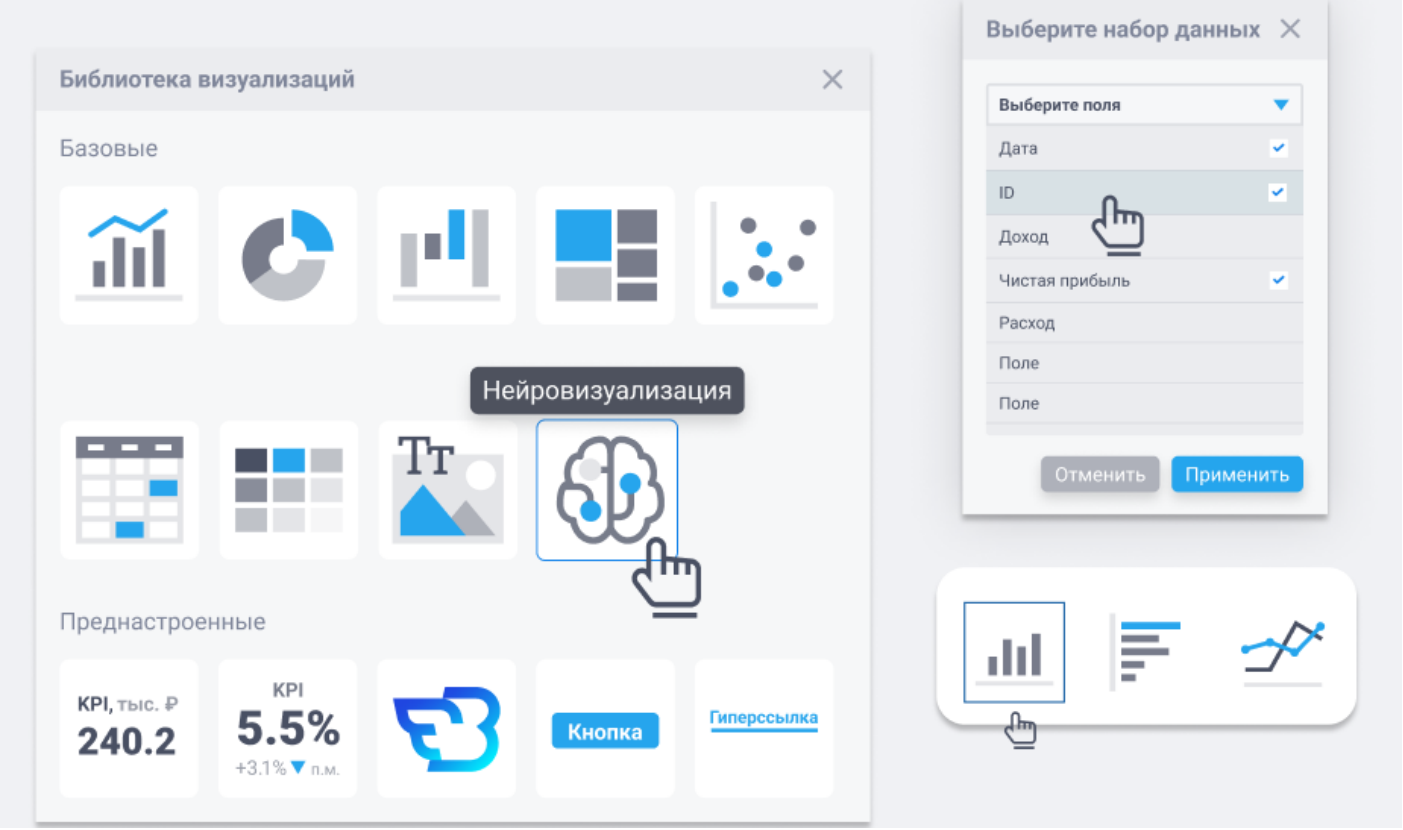
Нейровизуализации

Цель

Интеллектуальная система, которая на основе выбранных пользователем столбцов из модели данных генерирует несколько логически обоснованных визуализаций и предлагает их на выбор. Каждая визуализация должна сопровождаться подписью с указанием использованных типов полей и контекста, в котором она может использоваться (сравнение, распределение, состав или отношение)

Концепт для интерфейса



В список визуализаций, доступных для выбора на дашборде, необходимо добавить модуль **«Нейровизуализация»**. При нажатии на этот модуль в окне должен открываться список с доступными моделями данных на выбор и полями в этих моделях. Пользователь может выбрать до 4-х полей, после чего система автоматически предложит ему от 2 до 6 наиболее подходящих вариантов визуализации на основе предоставленных данных.

Концепт ML

В модели используются 4 расходящиеся ветки для определения контекста выбранных данных. Нужна модель ИИ или алгоритм с использованием разведочного анализа данных, позволяющие выполнить задачу классификации входных данных на наиболее подходящие из следующих классов задач: Сравнения, Распределения, Состава (Структуры) или Отношения. На входе модель получает от пользователя только столбцы данных (от 1 до 4) из итоговой таблицы. Возможно подключение вероятностной модели для определения наилучшего контекста.

После определения контекста задачи требуется применение алгоритма для выбора конкретной визуализации. На выходе модель должна предоставлять следующие данные:

- Наборы столбцов (от 2 до 6) с разделением на показатели (данные для отображения) и разрезы (поля для группировки). Для начала типом агрегации для передачи в модель для всех данных будет "Сумма"!
- Тип визуализации (диаграмма, водопад и т.д.) для каждого набора столбцов для отрисовки на фронте.
- Контекст использования одним словом (1 из 4 классов задач, определенных моделью) к каждому набору столбцов.

Задание на ML-разработку

Задача классификации

На вход модели поступают от 2 до 4 столбцов из модели данных. Тип данных у каждого столбца может быть одного из двух видов:

- **Числовой** – типы данных Int32, Int64 и т.д., Float, Decimal и иные целочисленные и вещественные типы данных, для которых определены арифметические (сложение, вычитание, умножение, возведение в степень) и логические (больше, меньше, равно, в промежутке) операции. В основном выступают в качестве показателей.
- **Категориальный** – строковые (String, FixedString), булевы типы данных, дата и время, списки, массивы, а если проще – все остальные, нечисловые типы данных. Не используются в вычислениях (за исключением даты и времени) в визуализациях, выступают чаще в качестве разрезов.

При необходимости модель может извлекать из данных статистическую информацию: математическое ожидание, дисперсию, отклонения и т.д.

Основная задача модели – на основании имеющихся данных вероятностно классифицировать наборы столбцов по следующим классам:

- **Сравнение** – пользователь смотрит на разницу (динамику) каждого показателя (числового типа) в разрезе категориальных данных (сравнение по разрезам). Пример: динамика цен на продукты по месяцам; разница в поголовье скота на разных фермах; уровень заработной платы по сотрудникам; динамика курса валют за последний год и т.д.
- **Отношение** – пользователь смотрит на различия между совокупностями показателей (данными числового типа) для одной категории (сравнение по показателям). Пример: показатели роста и веса у студентов; содержание белков, жиров и углеводов в молочных продуктах; соотношение память – количество ядер – диагональ для моделей ноутбуков; количество продаж и прирост выручки фирм в регионе и т.д.
- **Состав или структура** – пользователь смотрит на эффект (различия, прирост, долю) отдельных категорий по конкретному числовому показателю (возможно, составному). Пример: доли продаж дочерних компаний в общем объёме продаж холдинга; прибыль/убыток от продаж магазина по категориям товаров; площадь, занимаемая каждым жилым комплексом в доле от общей площади строительства по городам, регионам, странам и т.д.
- **Распределение** – пользователь смотрит на размещение категориальных признаков согласно их числовым показателям. Пример: карта конкурентов (разделение на 4 категории по числовым осям и распределение на ней компаний в зависимости от их показателей); распределение затрат на отделы компании; результаты соревнований по бегу для определения победителя.

> Распределение в целом очень похоже и на сравнение, и на отношение (даже по выбору визуализаций). Тем не менее, если неверно отнести набор данных не к этому классу, то дальнейшее применение алгоритма по подбору визуализации может дать негативный результат.
> В отличие от отношения для распределения важнее визуализировать то, к какой категории принадлежит какой показатель, чем то, как показатели соотносятся между собой (к какой группе отнести конкурента, а не какой конкурент доминирует над

> остальными); от сравнения – чем то, какая разница между показателями разных категорий (какому спортсмену выдать медаль, а не выявить отставание конкретного спортсмена от победителя).

Наборы столбцов необходимо собирать по всем доступным вариантам из тех, что предоставил пользователь. Если на входе модель получает:

- **2 столбца** от пользователя, то есть только **1 вариант** их сочетания;
- **3 столбца** от пользователя, то есть **4 варианта** их сочетания (3 пары из двух столбцов и 1 тройка);
- **4 столбца** от пользователя, то есть **11 вариантов** их сочетания (6 пар, 4 тройки и 1 четвёрка).

Для КАЖДОГО варианта сочетания столбцов модель должна рассмотреть ВСЕ 4 контекста использования (класса). В итоге получится **от 4 до 44 возможных выходных результатов** вида "набор данных – класс".

Далее необходимо расставить вероятности для каждого из выходных результатов. Для решения именно этой задачи требуется модель ИИ. Модель должна проанализировать метаданные каждого столбца из набора данных, рассчитать все необходимые дополнительные величины (например, статистические параметры) и определить, насколько каждый класс подходит каждому набору данных в процентном соотношении.

Важным требованием является наличие **границы отсечения** процентно-незначимых наборов выходных данных. Значение этой границы можно получить опытным путём или установить искусственно с возможностью дальнейшей корректировки в процессе использования модели (например, установить границу равной 85%). После получение процентного соответствия классов наборам столбцов возможен один из следующих результатов сравнения с границей:

- **Не более чем 1** набор данных прошёл границу отсечения. В таком случае принимаем удовлетворительными два набора данных с наибольшими значениями процентного соответствия выбранному классу.
- Границу отсечения прошли **от 2 до 6** (включительно) наборов данных. В таком случае принимаем удовлетворительными именно эти наборы данных.
- Границу отсечения прошли **более 6** наборов данных. В таком случае принимаем удовлетворительными шесть наборов данных с наибольшими значениями процентного соответствия выбранному классу.

Удовлетворительные наборы данных вместе с контекстом использования (классом) отправляются на вход алгоритма выбора визуализации.

Алгоритм выбора визуализации

* **Тип данных** читать как **столбец с типом данных** (н-р, числовой тип данных в показателе = столбец с числовым типом данных в показателях).

Представляет собой обычное дерево решений, первая ветка которого определяется классом набора данных. Дальнейший выбор зависит от типов данных набора и некоторых других характеристик, записанных условиями.

> До перехода на одну из веток дерева решений должна производиться проверка: если все выбранные пользователем столбцы имеют категориальный тип данных (не числовой), то вместо перехода на любую из веток эти данные собираются в одну таблицу (в разрезах) без других вариантов. Вместо класса ставится подпись: "Для получения корректного результата необходимо выбрать хотя бы одно числовое поле".

> Аналогичная проверка должна выполняться и для числовых типов данных, но с одним исключением: если хотя бы в одном столбце из набора не более 12 уникальных значений, то набор данных необходимо пропустить на одну из веток (этот числовой столбец будет играть роль категориального). Если таких столбцов нет, то все данные собираются в одну таблицу (в разрезах) без других вариантов. Вместо класса ставится подпись: "Для получения корректного результата необходимо выбрать хотя бы одно категориальное поле".

> Если на выходе модели представлено более одной таблицы, то из всех таблиц должна остаться только одна (с наибольшим числом столбцов), а остальные – быть удалены.
> *Если на выходе есть и таблицы, и другие визуализации, то из таблиц остаётся одна, а другие визуализации проходят без изменений

Сравнение:

1. Есть столбцы категориального типа И Все столбцы категориального типа имеют тип данных "дата (дата и время)" И Уникальных значений в категориальных типах НЕ более 12:
    1. Только один столбец числовой:
        - Визуализация **Комбинированная**; тип графика – **линия**, категориальные типы данных в разрезах, числовой тип данных в показателе (один линейный график).
    2. Несколько столбцов числовые (хотя бы один – категориальный):
        - Визуализация **Комбинированная**; тип графика – **линия**, категориальные типы данных в разрезах, числовые типы данных в показателях (несколько линейных графиков).
2. Нет столбцов категориального типа ИЛИ Есть столбцы категориального типа, отличающегося от типа данных "дата (дата и время)" ИЛИ Уникальных значений в категориальных типах более 12:
    1. Только один столбец числовой:

- Визуализация **Комбинированная**; тип графика – **столбик**, категориальные типы данных в разрезах, числовой тип данных в показателе (одна столбчатая визуализация).

2. Несколько столбцов числовые:
   - Визуализация **Комбинированная**; тип графика – **столбик**, категориальные типы данных в разрезах, числовые типы данных в показателях (несколько столбчатых визуализаций).
3. Все столбцы – числовые:
   1. Хотя бы в одном столбце НЕ более 12 уникальных значений:
      - Визуализация **Комбинированная**; тип графика – **столбик**, ОДИН числовой тип данных с наименьшим числом уникальных значений в разрезах, остальные числовые типы данных в показателях (одна или несколько столбчатых визуализаций).
   2. Во всех столбцах более 12 уникальных значений:
      - Визуализация **Таблица**; числовые типы данных в разрезах.

Отношение:

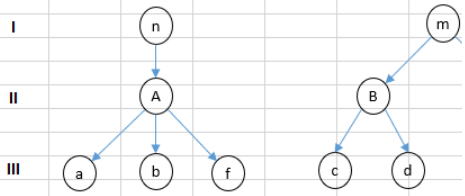
1. В наборе только один столбец с числовым типом данных ИЛИ более одного столбца с категориальным типом данных (при наличии хотя бы одного с числовым):
   - **СООБЩИТЬ ОБ ОШИБКЕ!** – модель неверно классифицировала контекст как Отношение. Исключить ветку из рассмотрения.
2. В наборе два столбца с числовым типом данных и один столбец с категориальным:
   - Визуализация **Пузырьковая**; категориальный тип данных – в разрезе, числовые типы данных – в показателях (без размера пузырьков).
3. В наборе три столбца с числовым типом данных и один столбец с категориальным:
   - Визуализация **Точечная**; категориальный тип данных – в разрезе, числовые типы данных – в показателях, один (случайным образом определенный) – в размере пузырька.
4. В наборе нет столбцов с категориальным типом данных (все столбцы – числовые):
   1. В наборе три столбца и хотя бы в одном столбце НЕ более 12 уникальных значений:
      - Визуализация **Пузырьковая**; ОДИН числовой тип данных с наименьшим числом уникальных значений в разрезах, остальные числовые типы данных в показателях (без размера пузырьков).
   2. В наборе четыре столбца и хотя бы в одном столбце НЕ более 12 уникальных значений:
      - Визуализация **Точечная**; ОДИН числовой тип данных с наименьшим числом уникальных значений в разрезах, остальные числовые типы данных в показателях, один (случайным образом определенный) – в размере пузырька.
   3. В наборе суммарно один или два столбца:
      - **СООБЩИТЬ ОБ ОШИБКЕ!** – модель неверно классифицировала контекст как Отношение. Исключить ветку из рассмотрения.

Состав или структура:

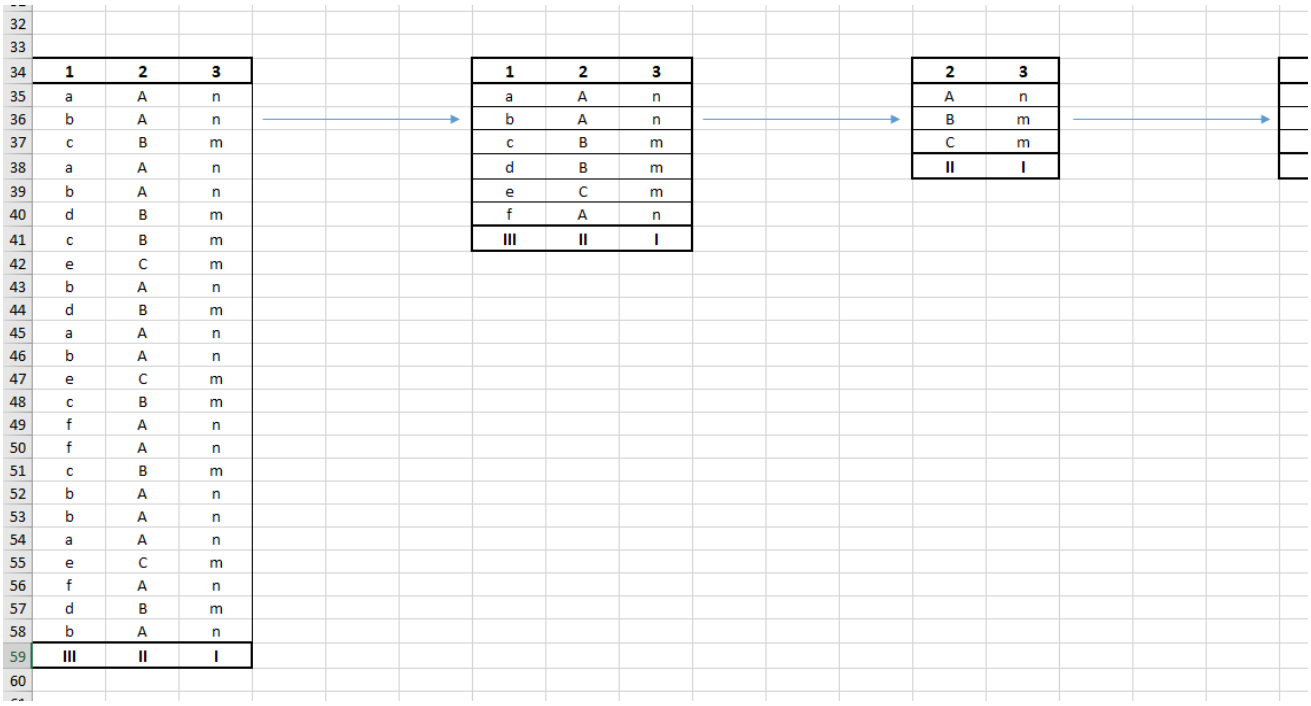
1. В наборе только один столбец с числовым типом данных:
    1. Алгоритмы на выявление **иерархических связей** в наборе данных, при которых каждому дочернему элементу соответствует только один родительский (можно использовать любой другой, вписывающийся в концепцию):
    Алгоритм 1: оставляем только категориальные столбцы, удаляем дубликаты строк (только для алгоритма), выполняем перебор всех возможных комбинаций категориальных столбцов. Если найдётся хотя бы одна комбинация, в которой каждому уникальному значению n-го столбца будет соответствовать только одно уникальное значение (n+1)-го столбца, то алгоритм выполнен.

1	2	3
a	A	n
b	A	n
c	B	m
a	A	n
b	A	n
d	B	m
c	B	m
e	C	m
b	A	n
d	B	m
a	A	n
b	A	n
e	C	m
c	B	m
f	A	n
f	A	n
c	B	m
b	A	n
b	A	n
a	A	n
e	C	m
f	A	n
d	B	m
b	A	n
III	II	I

1	2	3
a	A	n
a	A	n
a	A	n
a	A	n
b	A	n
b	A	n
b	A	n
b	A	n
b	A	n
b	A	n
f	A	n
f	A	n
f	A	n
c	B	m
c	B	m
c	B	m
c	B	m
d	B	m
d	B	m
d	B	m
e	C	m
e	C	m
e	C	m
III	II	I



    Алгоритм 2: оставляем только категориальные столбцы, удаляем дубликаты строк выполняем перебор всех возможных комбинаций категориальных столбцов. Проверяем первый столбец в комбинации. Если все значения в нём уникальные, то рассматриваем набор данных без этого столбца и переходим ко второму (по счёту в изначальном наборе) столбцу. Удаляем все дубликаты строк из таблицы, если все оставшиеся во втором столбце значения уникальные, то рассматриваем набор без этого столбца и переходим к третьему. Продолжаем до последнего столбца. Если удастся хотя бы в одной комбинации сократить выборку до последнего столбца, то алгоритм выполнен.



1	2	3
a	A	n
b	A	n
c	B	m
a	A	n
b	A	n
d	B	m
c	B	m
e	C	m
b	A	n
d	B	m
a	A	n
b	A	n
e	C	m
c	B	m
f	A	n
f	A	n
c	B	m
b	A	n
b	A	n
a	A	n
e	C	m
f	A	n
d	B	m
b	A	n
III	II	I

1	2	3
a	A	n
b	A	n
c	B	m
d	B	m
e	C	m
f	A	n
III	II	I

2	3
A	n
B	m
C	m
II	I

Если алгоритм выполнен полностью:

- Визуализация **Дерево**; категориальные типы данных в разрезах в порядке, обратном порядку в успешном алгоритме (сначала родительские, потом дочерние), числовые типы в показателях.

2. Если алгоритм не выполнен, то:
   - Визуализация **Круговая**; категориальные типы данных в разрезах, числовые типы в показателях.

2. В наборе РОВНО два числовых столбца и один категориальный:
   - Визуализация **Водопад**; категориальный тип данных в разрезе, если есть столбцы с названием "план" и/или "факт", то используем их в одноименных показателях, иначе: первый столбец в показателе "план", второй – в показателе "факт".
3. В наборе больше одного числового столбца И не выполнен пункт 2:
   1. Хотя бы в одном категориальном столбце столбце более 12 уникальных значений:
      - Визуализация **Комбинированная**; тип графика – **столбик**, объединить столбцы, категориальные типы данных в разрезах, числовые типы данных в показателях (накопительные столбчатые визуализации).
   2. Во всех столбцах более 12 уникальных значений:
      - Визуализация **Комбинированная**; тип графика – **линия**, установить непрозрачность области равной 50%, категориальные типы данных в разрезах, числовые типы данных в показателях (диаграмма области).

Распределение:

1. В наборе больше двух числовых столбцов:
   - **СООБЩИТЬ ОБ ОШИБКЕ!** – модель неверно классифицировала контекст как Распределение. Исключить ветку из рассмотрения.
2. В наборе один числовой столбец:

1. Хотя бы в одном категориальном столбце столбце более 12 уникальных значений:
    - Визуализация **Комбинированная**; тип графика – **столбик**, категориальные типы данных в разрезах, числовые типы данных в показателях (несколько столбчатых визуализаций).
2. Во всех столбцах более 12 уникальных значений:
    - Визуализация **Комбинированная**; тип графика – **линия**, установить толщину линии равной 0, толщину точки равной 6, категориальные типы данных в разрезах, числовые типы данных в показателях (точечный график).
3. В наборе два числовых столбца:
    - Визуализация **Пузырьковая**; категориальный тип данных – в разрезе, числовые типы данных – в показателях (без размера пузырьков).

Выход модели

Модель передаёт на фронт следующую информацию:

- Итоговые наборы данных, которые прошли границу отсечения классификатора и получили визуализацию;
- Контекст использования – один из четырёх присвоенных классов;
- Тип визуализации;
- Характеристики визуализации (распределение данных по разрезам и показателям, тип графика, толщина линии и т.д.)

Revision #1
Created 15 November 2024 05:16:44 by Артём
Updated 15 November 2024 05:16:44 by Артём