Лабораторная работа №1.

КЛАСТЕРИЗАЦИЯ. ПОСТРОЕНИЕ МОДЕЛИ КЛАСТЕРИЗАЦИИ В ORANGE

Цель и задача работы Изучить основные методы кластеризации с использованием приложения «Orange Data Mining».

Ход работы

Orange – свободно распространяемая библиотека, написанная на языке Python, основанная на принципе визуального программирования для наглядного доступа к алгоритмам Data mining. (http://orange.biolab.si/)

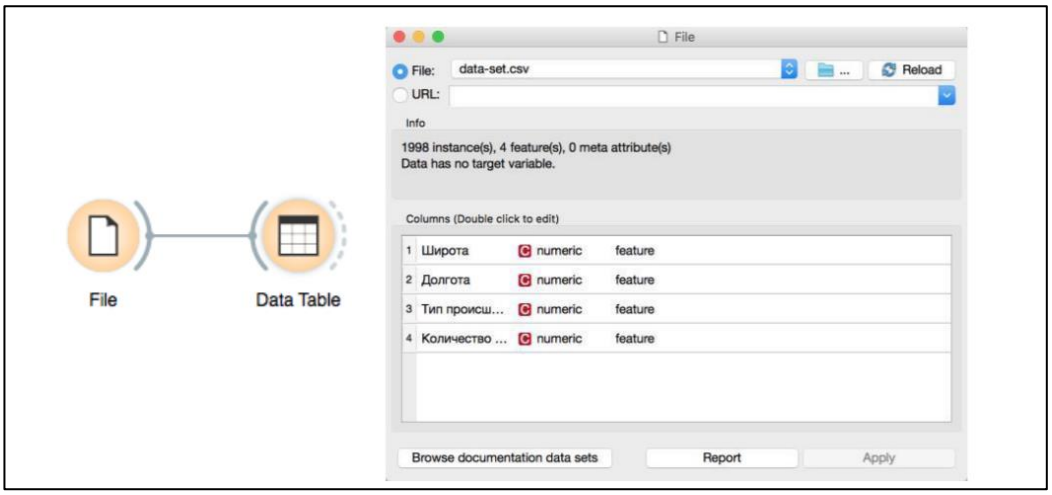
Orange предоставляет пользователю следующие
основные функции :

- Загрузка данных из различных источников (файлы, веб-ресурсы, базы данных) и представление их в табличном виде.
- Получение информации об атрибутах данных (полях таблицы).
- Построение потока data mining (data mining workflow).
- Изменение данных и параметров «на лету» (что позволяет отследить изменения в режиме реального времени).
- Визуализация результатов с помощью различных графиков.
- Сохранение модели и применение её в дальнейшем.

Пример построения модели Data mining в Orange
Загрузка данных (виджет File)

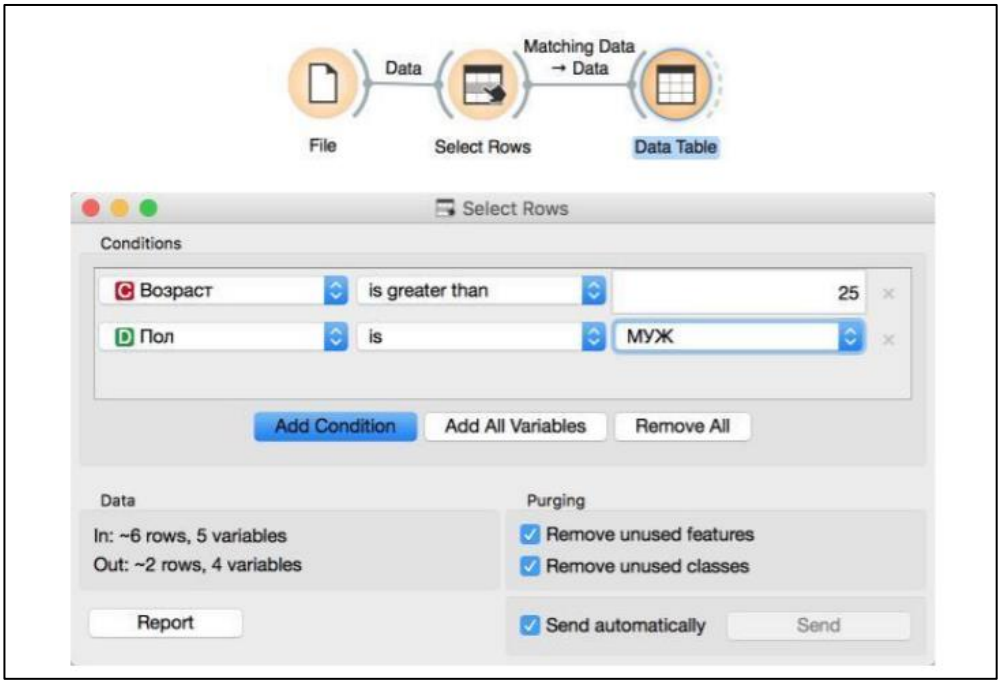
Orange поставляется со своим собственным форматом данных, но также может работать с другими форматами, например, Excel (.xlsx или .xls) или CSVфайлами. Как правило, входными данными является таблица с записями (объектами) в строках и атрибутами данных в столбцах. Атрибуты могут быть разного типа (непрерывные, дискретные и строковые). Типы атрибутов и их вид могут быть представлены в заголовке таблицы. Они также могут быть впоследствии изменены в виджете **File**. Тип данных также может быть изменен с помощью виджета **Select Columns.**

Виджет **File** находится на вкладке **Data.** Пример простой модели Orange с использованием виджетов File и Data Table показан на рис. 1



Очистка/фильтрация данных (виджеты Select Rows и Select Columns)

Для простой очистки данных по значению какого-либо атрибута можно использовать виджет **Select Rows**. Данный виджет позволяет задать различные условия для фильтрации входных данных. Например, у нас есть список пользователей и нам нужно исследовать только мужчин в возрасте от 25 лет и старше. Тогда настройки виджета Select Rows будут иметь вид, как показано на рис. 2.



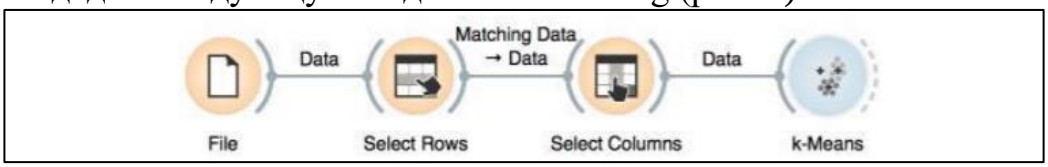
Допустим, что в приведенном выше примере нам нужно исследовать все поля, кроме поля ID. Для этого мы применим виджет **Select Columns** с настройками, показанными на рис. 3.

Можно самостоятельно придумать некоторый набор данных и посмотреть, как будут меняться выходные данные с помощью виджета **Data Table.** Данный виджет хорошо подходит для отображения табличных данных, полученных на выходе других виджетов.

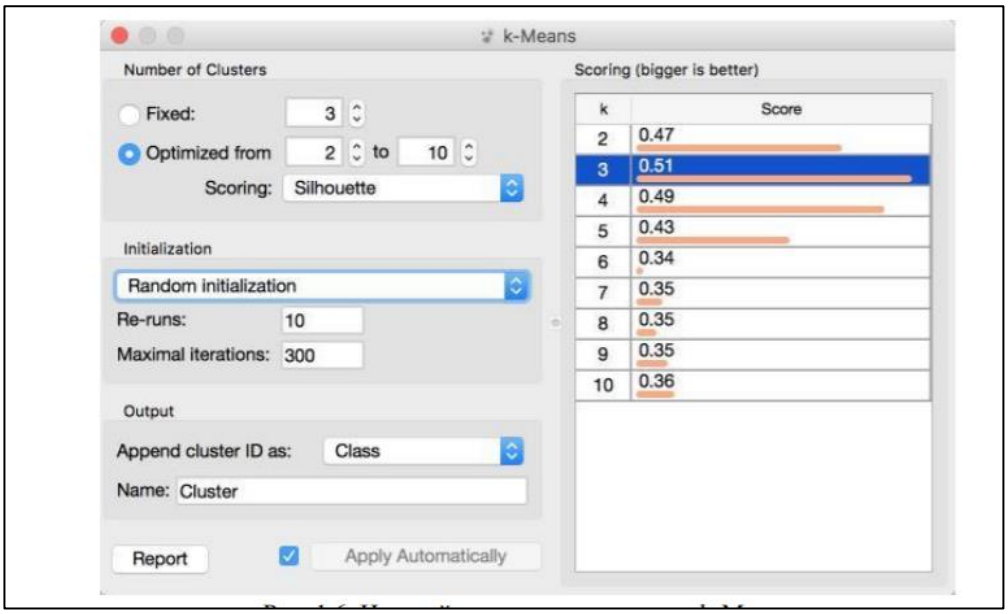
Все описанные виджеты также находятся на вкладке **Data.**

Использование виджета k-Means

Виджет алгоритма k-Means находится на вкладке **Unsupervised**. После необходимой подготовки данных мы может применить данный алгоритм. Для этого создадим следующую модель Data mining (рис. 4).



Остановимся подробнее на возможных настройках данного виджета. Так, мы можем устанавливать количество кластеров, на которые алгоритм пытается разделить наши данные (группа **Number of Clusters** на рис. 5).



Доступно два варианта: фиксированное (**Fixed**) количество кластеров и подобранное (**Optimized**). При выборе подбора количества кластеров необходимо задать минимальное и максимальное количество кластеров.

Также в блоке **Number of Cluster** можно выбрать критерий оценки качества кластеризации (Scoring) (рис. 5). Для выбора доступны следующие варианты:

– **Silhouette** – мера того, как сильно принадлежит объект своему кластеру и как сильно он отличается от объектов других кластеров. Другими словами, как среднее расстояние до объектов своего кластера соотносится со средним расстоянием до объектов других кластеров;

– **Inter-cluster distance** – мера расстояния между центрами кластеров;

– **Distance to centroids** – мера расстояния до объектов, соответствующих среднему арифметическому значению кластера.

В ходе лабораторной работы важно отметить, как влияет выбор критерия качества кластеризации на результат работы алгоритма.

Также на рис. 4 можно видеть блок **Initialization**. Он позволяет изменить алгоритм начального выбора центров кластеров. Доступны 2 значения:

1. **k-Means++** заключается в том, что первые центры кластеров выбираются случайно, последующие центры выбираются из оставшихся объектов с вероятностью, пропорциональной квадрату расстояния до ближайшего центра.

2. **Random initialization** – центры кластеров выбираются случайно и изменяются по ходу работы алгоритма.

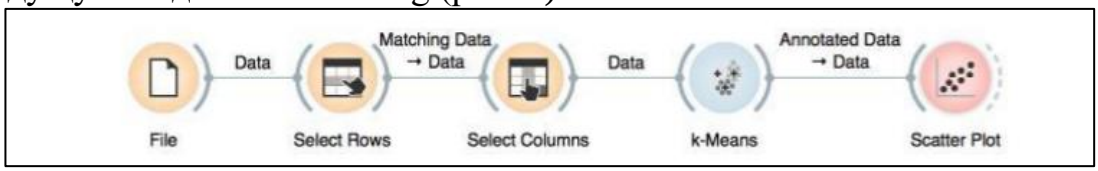
Визуализация с помощью виджета Scatter Plot

Виджет **Scatter Plot** находится на вкладке **Visualize**. Он позволяет строить диаграмму рассеивания (или точечную диаграмму). Добавим к предыдущему примеру модели data mining виджет **Scatter Plot** (рис. 6).

В настройках виджета можно выбирать оси координат (Axis x, Axis y), менять размер, цвет и вид маркеров (группа Points). Также имеется возможность сохранять полученную диаграмму в виде файла (кнопка Save Image).

Использование виджета Hierarchical Clustering

Виджет алгоритма иерархической кластеризации (**Hierarchical Clustering)** находится на вкладке **Unsupervised.** После необходимой подготовки данных можем применить данный алгоритм. Для этого дополним предыдущую модель data mining (рис. 7).



Алгоритм иерархической кластеризации не может работать напрямую с входными данными, для работы ему необходима матрица расстояний. Для этого добавим виджет **Distances**.

Рассмотрим возможные настройки виджета **Distances**

Существует возможность менять меру расстояния **(Distance Metric).** Доступны следующие варианты:

– **Euclidean** – эвклидово расстояние;

- **Manhattan** – также известное как расстояние городских кварталов: расстояние между двумя точками, равное сумме модулей разностей их координат;
- **Cosine** – косинусный коэффициент, или коэффициент сходства, определяется как косинус угла между двумя векторами внутреннего пространства объектов;
- **Jaccard** – мера Жаккара: размер пересечения делится на размер объединения множеств выборок; и другие.

Также настройки виджета **Hierarchical Clustering** позволяют задать способ выбора уровня кластеризации (блок **Selection**).

Результат работы виджета **Hierarchical Clustering** также можно отобразить на точечной диаграмме (**Scatter Plot**). Для этого нужно добавить соответствующий виджет и соединить его с виджетом **Hierarchical Clustering.**

Порядок выполнения работы

1) Загрузить исходные данные с помощью виджета File.
2) Осуществить необходимую фильтрацию/очистку данных с помощью виджетов Select Rows и Select Columns в соответствии с заданием.
3) Осуществить кластеризацию подготовленных данных с помощью алгоритма k-средних (k-means) с фиксированным и опциональным количеством кластеров.
4) Вывести результаты с помощью точечной диаграммы (виджет Scatter Plot) для всех экспериментов в ходе выполнения п. 3.
5) Осуществить кластеризацию подготовленных данных с помощью алгоритма иерархической кластеризации (виджет Hierarchical Clustering). Использовать в качестве меры расстояния евклидово, манхеттенское и любое другое расстояние на выбор.
6) Вывести результаты с помощью точечной диаграммы (виджет Scatter Plot) для всех экспериментов в ходе выполнения п. 5.
7) Сравнить результаты работы двух алгоритмов (k-средних и иерархической кластеризации). Сделать выводы.
8) Отключить фильтрацию данных и повторить эксперимент (только для оптимальных параметров). Сравнить полученные диаграммы с предыдущими.
9) Сделать выводы о влиянии фильтрации (очистки) данных на конечный результат. Подготовить отчёт.