МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**

**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ «МИСиС»**

*ИНСТИТУТ* ИТАСУ

*КАФЕДРА АВТОМАТИЗИРОВАННЫХ СИСТЕМ УПРАВЛЕНИЯ*

*НАПРАВЛЕНИЕ* 09.04.01 ИНФОРМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА

**Лабораторная работа №4**

**По дисциплине**: Технологии интеллектуального анализа данных

**На тему:** Применение технологий обучения без учителя и кластеризации на примере сегментации абонентов телекоммуникационной компании

**Студент: Костромин Д.А.**

**Группа: МИВТ 18-5-7**

**Преподаватель: Фомичева О.Е.**

**Оценка: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Дата защиты: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Москва, 2019

**Оглавление**

[**Введение** 3](#_Toc26877748)

[**1. Проблемы машинного обучения без учителя и способы оценки качества кластеризации** 4](#_Toc26877749)

[**2. Алгоритмы кластеризации** 7](#_Toc26877750)

[**2.1 Кластеризация k-средних** 7](#_Toc26877751)

[**2.2 Агломеративная кластеризация** 8](#_Toc26877752)

[**2.3 DBSCAN** 9](#_Toc26877753)

[**Выводы** 11](#_Toc26877754)

[**Список источников** 12](#_Toc26877755)

# **Введение**

К машинному обучению без учителя относятся все подходы, задачи и алгоритмы машинного обучения, когда отсутствует ответ (y) для работы алгоритма. В задачах такого типа алгоритму известны лишь входные данные, на основании которых он может найти некоторые закономерности. Основной проблемой в данном случае является проверка правильности работы алгоритмов, так как имеется лишь набор исходных данных без ответов, крайне сложно оценить корректность работы алгоритма.

В работе рассматриваются различные алгоритмы и особенности машинного обучения без учителя, включая снижение размерности исходных данных -- алгоритм t-SNE и преобразование N-мерного евклидова пространства признаков в евклидово пространство меньшего размера, например двумерное, которое легче визуализировать. И алгоритмы кластеризации – разделение исходных данных на несколько групп (кластеров), на основании их сходства. Кластеризация подобна классификации за тем исключением, что в случае кластеризации имеются лишь исходные данные, на основе которых алгоритм может предположить какие могут быть кластеры и как распределить данные между ними, но неизвестно сколько на самом деле кластеров. В работе рассматриваются такие алгоритмы кластеризации как Кластеризация k-средних (k-means), агломеративная кластеризация и DBSCAN.

В качестве примера рассматриваются сегментация абонентской базе телекоммуникационной компании, это классическая задача кластеризации без знания истинной кластеризации – реальных сегментов пользователей.

При выполнении работы будут использованы модели машинного обучения из библиотеки scikit-learn для языка Python.

# **1. Проблемы машинного обучения без учителя и способы оценки качества кластеризации**

Как уже упомянуто выше, одной из основных проблем обучения без учителя в целом и кластеризации в частности является затруднительное определение качества работы моделей. При этом существуют некоторые способы оценки качества кластеризации.

К ним относятся скорректированный коэффициент Рэнда (adjusted Rand index, ARI), принимающий значение от -1 до 1 и нормализованная взаимная информация (normalized mutual information, NMI), принимающая значение от 0 до 1, соответственно, чем ближе оценка данных метрик к единице, тем более достоверными являются результаты кластеризации. Но, ARI и NMI требуют знания истинной кластеризации – должны быть известны реальные метки классов кластеров, что делает их малоэффективными для большинства задач, включая рассматриваемую в данной работе, так как если уже известны имеющиеся в данных классы (например, известны возможные сегменты абонентов телекоммуникационной компании) то можно воспользоваться алгоритмами классификации и необходимость в обучении без учителя отпадает. С другой стороны, ARI и NMI нельзя назвать бесполезными, в частности их можно использовать чтобы узнать, насколько отличаются результаты кластеризации с использованием разных алгоритмов.

Существуют метрики оценки кластеризации, не требующие знания реальных кластеров, к ним относят силуэтный коэффициент (silhouette coefficient) – числовое значение от 0 до 1 отражающее компактность кластера, высокое значение которой указывает на сходство отнесенных к одному кластеру объектов, соответственно низкое на существенные отличия между ними. Но данная метрика малопригодна для реальных задач, в том числе, потому что “родственные” объекты могут существенно отличаться друг от друга, а силуэтных коэффициент в первую очередь ориентируется на компактность кластеров, “кучность” отражающих объекты точек в евклидовом пространстве. Если в качестве примера рассмотреть разделение на два кластера набора данных о животных, где имеются сведения о, например кошках и собаках, то рассматриваемая метрика скорее всего покажет низкий результат при близкой к достоверной кластеризации, так как различные породы собак могут сильно отличаться друг от друга ростом, размером и весом и как следствие отражающие их точки в евклидовом пространстве могут принять “некомпактную”, сложную структуру [1].

Еще одним способом оценки кластеризации является визуализация результатов, но здесь возникает серьезная проблема представления N-мерных данные в виде понятного человеку двумерного изображения. Одним из возможные решений является применение другого семейства алгоритмов машинного обучения без учителя – снижение размерности исходных данных и алгоритмы множественного обучения. В качестве примера множественного обучения можно привести алгоритм t-SNE, основная идея данного подходя заключается в преобразовании представления N-мерных исходных данных к более подходящему для решения задачи, это может быть сокращение чрезмерно большого числа признаков или визуализация многомерных данных в виде двумерных точечных диаграмм. При этом алгоритм t-SNE преобразует N мерные данные таким образом, чтобы после приведения к двумерному виду расстояние между характеризующими объекты точками в евклидовом пространстве было как можно ближе к исходному [1].

Стоит отметить, что алгоритмы кластеризации работают наилучшим образом, когда в исходных данных имеется несколько группы объектов, существенно отличающихся друг от друга, такие объекты можно легко отнести к различным кластерам, Рисунок 1.

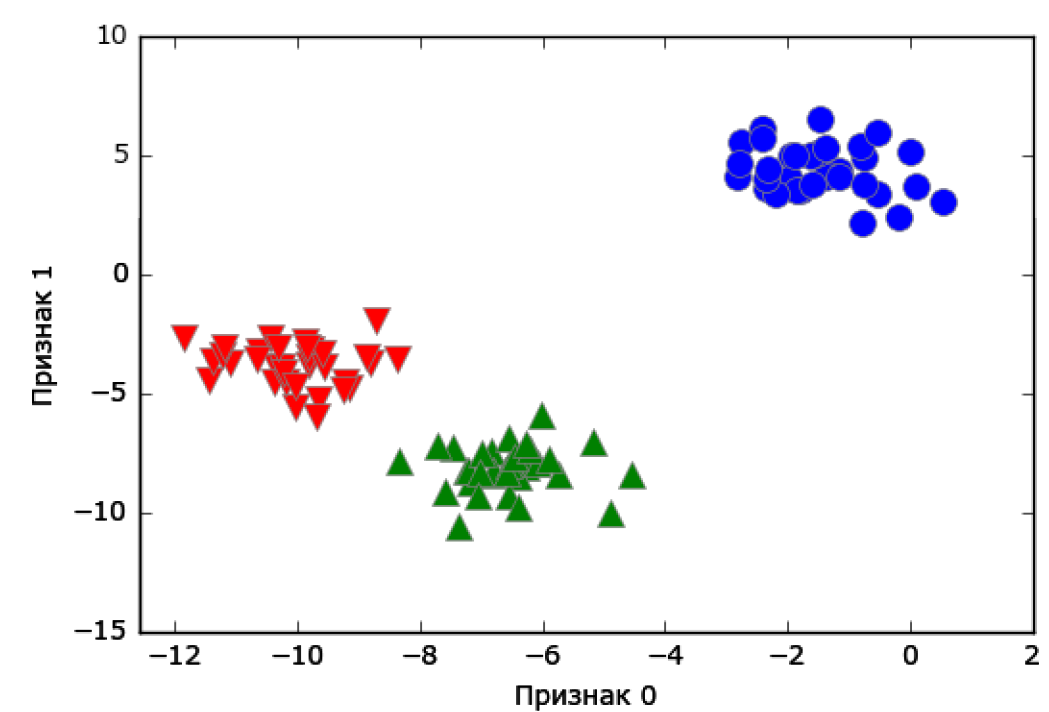


Рисунок 1Три кластера с существенными отличиями друг от друга

Ситуация усложняется, когда объекты сложно разделить. К сожалению, подобную ситуацию демонстрирует двумерное представление данных телекоммуникационной компании, по которому сложны выделить возможные кластеры и их количество, Рисунок 2

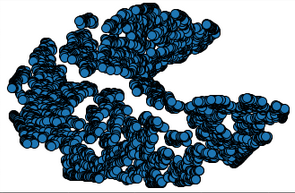


Рисунок 2 Данные о абонентах телекоммуникационной компании представленные в двумерном виде с помощью алгоритма t-SNE

# **2. Алгоритмы кластеризации**

Рассматриваемый набор данных о клиентах телекоммуникационной компании содержит 10 признаков (за исключением поля “Код”, не являющегося признаком), что существенно затрудняет визуализацию распределения данных, Рисунок 3

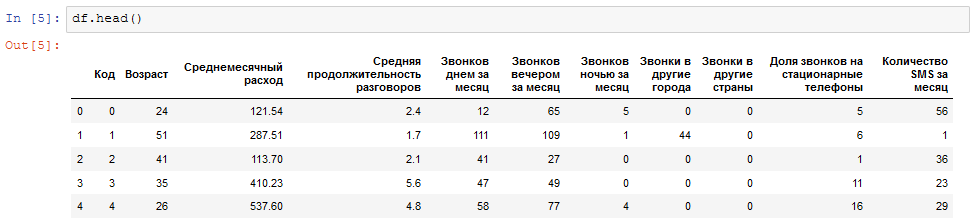


Рисунок 3 Данные о абонентах телекоммуникационной компании

С целью лучшей визуализации распределения данных по кластерам данные приведены к двумерному представлению с помощью алгоритма t-SNE, и представляют собой двумерный массив, который будет использован в дальнейшем для формирования и визуализации кластеров, Рисунок 4

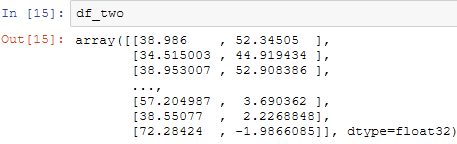


Рисунок 4 Данные о абонентах телекоммуникационной компании приведенные к двумерному представлению

## **2.1 Кластеризация k-средних**

Кластеризация k-средних является одним из наиболее простых алгоритмов кластеризации. Во время работы данного алгоритма сначала задается число кластеров, затем для каждого из них определяется объект обучающей выборки (точка в евклидовом пространстве), определяющая центр кластера. На следующем этапе для всех остальных объектов вычисляется расстояние в евклидовом пространстве до каждого из центров кластеров и соответственно объект относится к кластеру, расстояние между центром которого и объектом наименьшее.

К недостаткам алгоритма k-средних относится невозможность определения сложных кластеров – даже если известно сколько кластеров представлены в обучающей выборке, так как k-means предполагает, что все кластеры обладают одинаковым диаметром в евклидовом пространстве, что далеко не всегда встречается в реальности.

Можно продемонстрировать работу алгоритма k-средних на примере данных о абонентах телекоммуникационной компании, приведенных к двумерному представлению для лучшей визуализации с помощью алгоритма t-SNE, Рисунок 5.

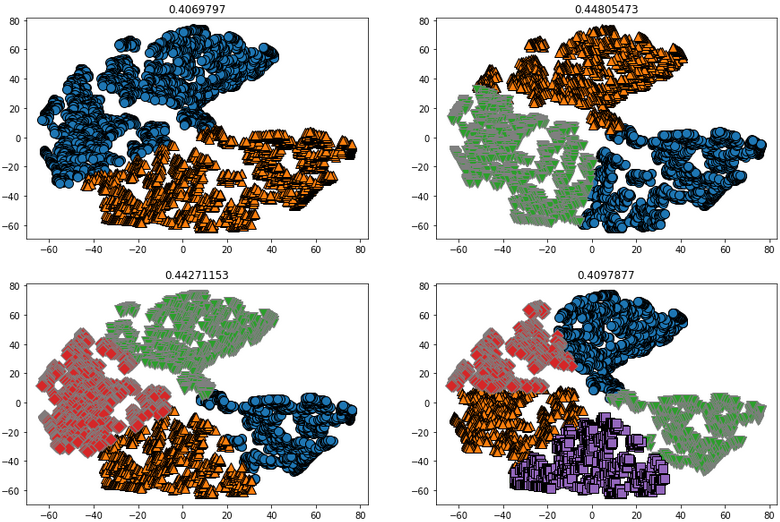


Рисунок 5 Кластеризации k-средних для 2-5 кластеров

## **2.2 Агломеративная кластеризация**

Другим алгоритмом кластеризации является агломеративная кластеризация. Идея данного алгоритма несколько отличается от k-средних, начиная работу алгоритм агломеративной кластеризации заносит каждый объект обучающей выборки в свой собственный кластер, затем начинает объединять наиболее похожие друг на друга кластеры до тех пор, пока не сработает определенный критерий остановки. Критерии остановки могут быть заданы различным образом, например для реализации данного алгоритма в библиотеке scikit learn критерием остановки является количество кластеров, как только формируется заданное перед началом работы алгоритма число кластеров, он завершает работу. В отличии от алгоритма k-means агломеративная кластеризация формирует кластеры различные по форме и диаметру в евклидовом пространстве, что позволяется искать закономерности в сложных структурах данных.

По аналогии с k-средних можно продемонстрировать работу агломеративной кластеризации на данных о телекоммуникационной компании, рисунок 6.

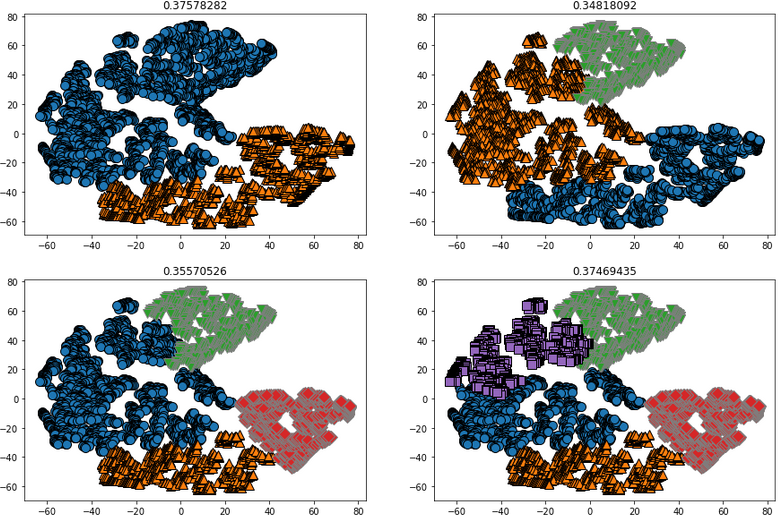


Рисунок 6 Агломеративная кластеризация для 2-5 кластеров

## **2.3 DBSCAN**

Принципиально другой подход к кластеризации предлагает алгоритм DBSCAN (плотностный алгоритм кластеризации пространственных данных с присутствием шума) в отличии от ранее рассмотренных алгоритмов DBSCAN самостоятельно определяет количество кластеров в выборке, не требую явно указывать их количество перед началом работы. Также данных алгоритм способен определять “шум” – объекты, которые невозможно отнести ни к одному из кластеров. Подобных подход с одной стороны позволяет осуществлять кластеризацию данных со сложной структурой, но с другой стороны делает данных алгоритм сложным с точки зрения настроек и интерпретации результатов. К тому же DBSCAN существенно более требователен к вычислительным ресурсам компьютера.

Можно привести пример кластеризации абонентов телекоммуникационной компании с помощью алгоритма, но если отправить на вход алгоритма двумерное представление исходных данных, то все объекты будут отнесены к одному кластеру, точнее – помечены как шум, рисунок 7.

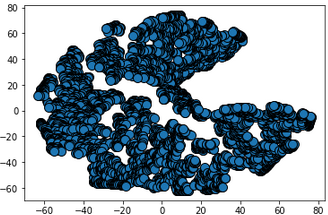


Рисунок 7 DBSCAN и обработка исходных данных

DBSCAN чувствителен к представлению исходных данных, можно преобразовать данные таким образом чтобы признаки объектов имели среднее значение равное нулю и дисперсию равную единице. Преобразование осуществляется для двумерного представления исходных данных, Рисунок 8

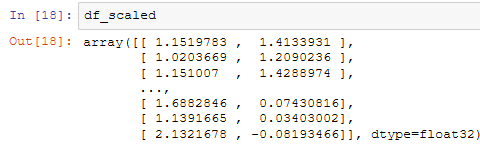


Рисунок 8 Масштабирование двумерного представления данных

Теперь можно повторно применить алгоритм DBSCAN для отмасштабированных данных, в таком случае результат его работы будет заметно отличаться – абоненты будет распределены между шестью кластерами, имеющими различную размерность, при этом ни один из объектов не будет помечен как шум, Рисунок 9.

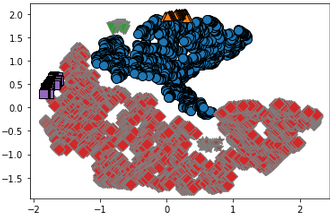


Рисунок 9 DBSCAN и обработка данных после масштабирования (StandartScaler) данных

# **Выводы**

Машинное обучение без учителя является отдельным и весьма сложным разделом машинного обучения. Главной особенностью таких алгоритмов является наличие исключительно входных данных, на основе которых модель должно построить закономерности, но не ответов. Что существенным образом затрудняет применение и интерпретацию результатов работы алгоритмов.

Существуют различные подтипы алгоритмов машинного обучения без учителя, к ним относятся кластеризация, масштабирование и изменение размерности исходных данных, а также алгоритмы извлечения ассоциативных правил. Стоит отметить, что в некоторых случаях машинное обучение без учителя применяется и в классических задачах обучения с учителем, например с целью предварительной обработки исходных данных, в частности для масштабирования и/или изменения размерности данных.

В работе были рассмотрены три различных алгоритма кластеризации данных на примере сегментации абонентов телекоммуникационной компании, результаты работы алгоритмов заметны отличаются друг от друга, при этом затруднительно сделать выводы о эффективности их работы как в целом, так и в сравнении друг с другом. Например, k-средних показывает большее значение метрики “силуэтный коэффициент” в сравнении с агломеративной кластеризацией, но из-за специфики работы силуэтного коэффициента сложно сказать, указывает ли это на превосходство k-means над агломеративной кластеризацией.

# **Список источников**

1. Андреас Мюллер, Сара Гвидо. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными. — СПб.: Диалектика, 2017. — 480 с.: ил.
2. Официальная документация библиотеки scikit learn, URL: -- https://scikit-learn.org/stable/modules/classes.html#module-sklearn.cluster