

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**

Кафедра математической кибернетики и компьютерных наук

**ИССЛЕДОВАНИЕ И МОДИФИКАЦИЯ АЛГОРИТМА
КЛАСТЕРИЗАЦИИ ДАННЫХ K -МХТ.**

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студентки 4 курса 411 группы
направления 02.03.02 — Фундаментальная информатика и информационные
технологии
факультета КНиИТ
Степановой Анастасии Александровны

Научный руководитель
доцент, к. ф.-м. н.

Ю. Н. Кондратова

Заведующий кафедрой
к. ф.-м. н.

С. В. Миронов

Саратов 2019

СОДЕРЖАНИЕ

ВВЕДЕНИЕ	3
1 Описания алгоритмов и способы оценки качества кластеризации	5
2 Описание программного кода	7
3 Результаты работы алгоритмов	9
ЗАКЛЮЧЕНИЕ	10
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ	12

ВВЕДЕНИЕ

В настоящий момент алгоритмы кластеризации данных применяется при решении многих задач. Существует большое количество алгоритмов кластеризации. Алгоритм k -МХТ был предложен в работе [1] для решения задачи кластеризации геотегированных данных. В работе [2] на основе фотографий с геотегами, полученных с сайта Flickr, выделялись популярные места города Саратова с помощью алгоритмов кластеризации DBSCAN и k -МХТ.

В задачах кластеризации данных с геотегами необходимо разделить на кластеры точки, лежащие в двухмерном пространстве поверхности шара. В настоящее время существует множество алгоритмов кластеризации геоданных. Каждый из них имеет свои особенности поведения на различных наборах данных. Например алгоритм k -Means [3–6] плохо кластеризует данные, имеющие вид ленты — данные, где для любой точки на близком расстоянии найдется другая точка, принадлежащая иному кластеру, а также существуют точки, которые не располагаются на близком расстоянии, но принадлежат одному и тому же кластеру. Алгоритм DBSCAN наоборот хорошо работает на данных ленточного вида, в то же время алгоритм DBSCAN не может правильно разбивать на кластеры данные, имеющие большую разницу в плотности.

В данной работе будет рассмотрена кластеризация данных с геоинформацией (геотегами). Для данных с геотегами мерой схожести объектов (точек) является расстояние между этими точками.

Целью бакалаврской работы является изучение алгоритмов k -Means и k -МХТ, поиск возможности улучшения алгоритма k -МХТ. Для этого необходимо:

1. изучить алгоритмы кластеризации k -МХТ и k -Means;
2. исследовать поведение алгоритма k -МХТ на различных типах данных, найти возможности его улучшения, сформулировать новый алгоритм кластеризации;
3. подготовить данные для кластеризации;
4. определить метрики оценки качества кластеризации;
5. реализовать заданные алгоритмы кластеризации, получить результаты работы этих алгоритмов на подготовленных данных;
6. провести анализ результатов работы алгоритмов кластеризации, на ос-

нове этого сделать выводы;

7. применить алгоритм кластеризации k -МХТ и его улучшенную модификацию в практической задаче выделения популярных мест Санкт-Петербурга на основе фотографий с геотегами, с помощью метрик оценки качества кластеризации сравнить качество кластеризации.

Теоретическая значимость данной работы заключается в том, что была создана модификация алгоритма кластеризации данных k -МХТ — алгоритм k -МХТ-Gauss. Алгоритмы кластеризации адаптированы под конкретную специфику наборов данных. Алгоритм k -МХТ-Gauss лучшим образом разбивает на кластеры наборы точек в декартовой системе координат и наборы точек на поверхности Земли, в отличие от алгоритма k -МХТ, что подтверждают рассмотренные в данной работе результаты кластеризации на смоделированных данных и значения метрики оценки качества кластеризации.

Практическая значимость в данной работе заключается в том, что модификацию алгоритма кластеризации данных k -МХТ — алгоритм кластеризации k -МХТ-Gauss, можно успешно применять для кластеризации данных с геотегами. В работе проводилось сравнение работы алгоритмов k -МХТ, k -МХТ-Gauss на задаче выделения популярных объектов города Санкт-Петербурга по фотографиям с геотегами. Данные с геотегами были получены с сайта Flickr. При сравнении результатов работы этих алгоритмов кластеризации было видно, что алгоритм k -МХТ-Gauss производит разбиение вершин на кластеры более правильным образом, что подтверждают метрики оценки качества кластеризации.

Бакалаврская работа состоит из введения, 3 разделов, заключения, списка использованных источников и 8 приложений. Общий объем работы — 66 страниц, из них 40 страниц — основное содержание, включая 22 рисунков и 6 таблиц, список использованных источников информации — 32 наименований.

1 Описания алгоритмов и способы оценки качества кластеризации

Данная глава описывает алгоритмы кластеризации данных k -МХТ, k -Means. В этой главе была предложена модификация алгоритма k -МХТ — алгоритм кластеризации данных k -МХТ-Gauss.

В алгоритме k -МХТ-Gauss необходимо построить граф G , вершинами которого являются точки на плоскости. Дугами графа G являются связи между вершинами, расположенные на расстоянии не превосходящим ε .

Необходимо посчитать вес всех дуг и оставить из каждой вершины k исходящих дуг с максимальным весом. Если необходимо выбрать из нескольких дуг с одинаковым весом, то эти дуги выбираются случайным образом.

Вес дуги равен количеству общих соседей соединяемых вершин умноженному на значение функции плотности распределения Гаусса (1) от расстояния между соединяемыми вершинами.

Поскольку была использована плотность распределения Гаусса, по правилу трех сигма для полного покрытия всех точек, расположенных на расстоянии ε от заданной вершины, достаточно взять $\sigma = \frac{\varepsilon}{3}$.

$$N(x, \sigma, \mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (1)$$

Функция плотности распределения Гаусса, где $\mu = 0$, $\sigma = \frac{\varepsilon}{3}$;

Кластерами будут являться компоненты сильной связности полученного графа.

В алгоритме k -МХТ при подсчете весов дуг в явном виде не используется расстояние между двумя вершинами. Алгоритм k -МХТ-Gauss отличается от алгоритма k -МХТ способом подсчета весов дуг. В алгоритме k -МХТ вес дуги равен количеству общих соседей соединяемых вершин.

Так же были описаны метрики оценки качества кластеризации ARI (Adjusted Rand Index) и Modularity, которые используются при сравнение результатов работы алгоритмов.

Метрика ARI (Adjusted Rand Index) сравнивает на сколько схожи два разбиения вершин на кластеры.

Modularity является одной из самых распространенных метрик оценки качества кластеризации, не требующей знания истинной кластеризации. Данная метрика основана на утверждении, что кластерами являются некоторые

области наибольшей плотности графа.

Таким образом в этой главе был подробно представлен теоретический материал, используемый в работе.

2 Описание программного кода

Программный код реализован на языке Python.

Самостоятельно были реализованы алгоритмы кластеризации k -МХТ, k -МХТ-Gauss. Так же самостоятельно был реализован подсчет значения метрики Modularity. С помощью библиотеки scikit-learn были сформированы наборы данных для проведения сравнения результатов работы алгоритмов.

Данные типа blobs были получены с помощью функции `sklearn.datasets.make_blobs` со следующими параметрами:

- `n_samples` — общее количество генерируемых точек
- `cluster_std` — стандартные отклонения кластеров
- `random_state` — параметр, управляющий генератором случайных чисел

Данные типа circles были получены с помощью функции `sklearn.datasets.make_circles` со следующими параметрами:

- `n_samples` — общее количество генерируемых точек
- `noise` — добавление к данным Гауссовского шума со стандартным отклонением
- `factor` — коэффициент масштабирования между внутренним и внешним кругом

Данные типа moons были получены с помощью функции `sklearn.datasets.make_moons` со следующими параметрами:

- `n_samples` — общее количество генерируемых точек
- `noise` — добавление к данным Гауссовского шума со стандартным отклонением

В работе использовались наборы данных полученные при различных значениях параметров функции

`sklearn.datasets.make_blobs` для получения наборов данных с одинаковой плотностью кластеров и с разной плотностью кластеров.

С помощью API сайта Flickr были получены данные с геотегами для города Санкт-Петербург.

В работе было оптимизировано время работы алгоритмов кластеризации с помощью левого и правого бинарных поисков.

Таким образом в работе была описана структура программного кода необходимого для получения набора данных, алгоритмов кластеризации дан-

ных и проведения сравнения рассмотренных в работе алгоритмов и представлен способ оптимизации времени работы алгоритмов кластеризации k -МХТ, k -МХТ-Gauss.

3 Результаты работы алгоритмов

Было произведено сравнение результатов работы алгоритмов k -МХТ, k -МХТ-Gauss, k -Means на смоделированных данных, а так же сравнение результатов работы алгоритмов кластеризации k -МХТ и k -МХТ-Gauss при выделение популярных объектов города Санкт-Петербург.

В результате проведенных исследований можно сделать выводы:

1. На смоделированных данных алгоритм k -МХТ-Gauss имеет большой диапазон значений ε , при которых метрика имеет максимальное значение.
2. Алгоритм k -МХТ-Gauss производит правильную кластеризацию данных типа blobs и circles.
3. Алгоритм k -МХТ-Gauss более производит правильную кластеризацию набора данных с геотегами для города Санкт-Петербург и достигает более высокого значения метрики Modularity, в отличие от алгоритма k -МХТ.

Полученные результаты были описаны в статье [7].

Таким образом в работе доказывается, что алгоритм k -МХТ-Gauss, лучше справляется с кластеризацией точек на кластеры, чем алгоритмы k -МХТ, k -Means.

ЗАКЛЮЧЕНИЕ

На смоделированных данных алгоритм k -МХТ-Gauss имеет большой диапазон значений ε , при которых $ARI = 1$, чего не наблюдается у алгоритма k -МХТ. Это позволит использовать алгоритм k -МХТ-Gauss для кластеризации данных в автоматическом режиме. Алгоритм k -МХТ-Gauss достигает значение метрики $ARI = 1$ при меньших значениях параметра k , чем алгоритм k -МХТ, что в некоторых практических задачах может существенно сократить время работы программы. Алгоритм k -МХТ-Gauss в отличие от алгоритма k -МХТ производит правильную кластеризацию данных, у которых граф имеет области с различной плотностью. Что позволяет успешно применять алгоритм k -МХТ-Gauss для задач кластеризации геоданных.

Алгоритм k -МХТ-Gauss в отличие от алгоритма k -Means правильным образом производит кластеризацию данных типа ленты.

Сравнение результатов работы алгоритмов k -МХТ и k -МХТ-Gauss на задаче выделения популярных мест города Санкт-Петербурга с помощью метрики Modularity подтверждает тот факт, что алгоритм k -МХТ-Gauss лучше справляется с кластеризацией данных в практических задачах, чем алгоритм k -МХТ. Алгоритм k -МХТ-Gauss в отличие от алгоритма k -МХТ при большом скоплении точек в одном месте не выделяет большое множество кластеров и не относит многие вершины к шумам, т. е. производит кластеризацию более качественно.

Сокращение диапазона поиска соседей каждой вершины с помощью левого и правого бинарного поиска существенно уменьшает время работы программы в практических задачах, за исключением случаев, где для каждой вершины графа соседями будут являться почти все вершины графа, т. е. в случаях, если граф имеет почти максимально возможное число дуг.

Алгоритм k -МХТ-Gauss можно успешно применять не только в задачах кластеризации геоданных, но и на любых наборах данных, где похожесть любых двух объектов может быть представлена с помощью числа, в таком случае параметр ε будет обозначать допустимую меру похожести между объектами.

В алгоритме k -МХТ-Gauss в зависимости от задачи можно вместо функции плотности распределения Гаусса при подсчете весов дуг использовать любую симметричную относительно нуля функцию с максимальным значением

в нуле. Что позволяет получать более правильное разбиение на кластеры данных с различной спецификой.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

- 1 *Cooper, C.* An experimental study of the k-mxt algorithm with applications to clustering geo-tagged data / C. Cooper, N. Vu. — 2018. — Vol. 10836. — Pp. 145–169. — doi: https://doi.org/10.1007/978-3-319-92871-5_10.
- 2 *Степанова, А.* Кластеризация геотегированных данных для поиска туристических достопримечательностей / А. Степанова, С. В. Миронов, Е. Коробов // *Математическое моделирование, компьютерный и натурный эксперимент в естественных науках*. — 2018. — № 3. — URL: <http://mathmod.esrae.ru/19-71> (Дата обращения 05.06.2019).
- 3 *Steinhaus, H.* Sur la division des corps materiels en parties / H. Steinhaus. — 1956. — Vol. 4. — Pp. 801–804.
- 4 *Lloyd, S.* Least squares quantization in pcm's. / S. Lloyd // *Bell Telephone Laboratories Paper*. — 1982. — Vol. 28, no. 2. — Pp. 129–137.
- 5 *Macqueen, J.* Some methods for classification and analysis of multivariate observations // In 5-th Berkeley Symposium on Mathematical Statistics and Probability. — 1967. — Pp. 281–297.
- 6 *Bottou, L.* Convergence properties of the k-means algorithms / L. Bottou, Y. Bengio. — MIT Press, 1995. — Pp. 585–592.
- 7 *Stepanova, A.* The clusterization of geo-tagged data for finding city sights with use of a modification of k-mxt algorithm / A. Stepanova, S. Mironov, E. Korobov, S. Sidorov // *Atlantis Press*. — 2019. — Vol. 85. — doi: <https://doi.org/10.2991/cmdm-18.2019.4>.