**Кластеризация с пустыми кластерами.**

1. **Аннотация**

Кластерный анализ данных применятся во многих прикладных задачах машинного обучения и анализа данных: медицине, социологии, экономике, кибербезопасности. Отсутствие наблюдений – не всегда отсутствие информации, поэтому предполагается, что наличие пустых кластеров может рассказать нам о наших данных, как и реальные наблюдения. Например, в задачах кредитного риска неправильное исходное определение кластера заёмщика может привести к ошибке оценки вероятности дефолта заёмщика. Некоторые прикладные исследователи предлагают использовать метку кластера как дополнительный признак в задаче регрессии (Kaggle.com [сайт], 2020), таким образом, ошибка в кластеризации может привести к смещению оценки регрессии.

Исследование пустых кластеров уже неявно рассматривалась в литературе, хотя и с определенными ограничениями. Цель работы состоит в том, чтобы начать с базового одномерного случая и предложить алгоритм, который выполняет одномерную кластеризацию для заданного числа кластеров и демонстрирует область из пустого кластера. Предложен алгоритм по определению потенциальных пустых кластеров и их размеров в зависимости от изначального разбиения выборки на количество кластеров. Реализован метод для заполнения этих пропусков и оценке смещения центроидов исходной кластеризации при учете пустого кластера.

1. **Степень научной разработанности темы, основные источники.**

В обзорной статье (Xu D., Tian Y. A., 2015) рассматривают актуальные SOTA подходы к кластерному анализу. Более классические алгоритмы рассматривают (Raschka, S. и Mirjalili, 2015) в 11 главе книги о машинном обучении. Пропуски в данных могут иметь разную природу и зачастую относятся к объясняющим переменным в задачах классификации или регрессии, при кластеризации же мы предполагаем, что мы не наблюдаем не только переменную, но целый кластер сущностей. Таким образом мы подразумеваем, что пропуск - не факт отсутствия кластера объектов как такового, а потенциально наличествующие объекты, отсутствующие в нашей выборке. Пустые кластеры можно объяснить по-разному, в зависимости от рассматриваемой задачи.

Зачастую, инициализация пустого кластера – это особенность реализации исходного алгоритма и этот случай рассматривается как недостаток метода, например, k-means, и такие случаи стараются минимизировать за счет модернизации классических алгоритмов. Данная ситуация может возникать, когда не соблюдены корректные начальные условия и оптимизатор находит локальный минимум, или же когда входные данные представлены бинарными или категориальными признаками (Raykov Y. P. и др., 2015). (Yadav A., Dhingra S., 2014, Pakhira M., 2009) предлагает модификацию алгоритма k-средних, который не инициализируют пустые кластера, путем добавления в расчёт положения центроидов gap-статистики или используя более современные оптимизаторы. (Hua C. и др., 2019) также предлагают альтернативу методу k-средних, генетический XK-Means, который также не инициализирует пустые кластеры. (Tavallali, P., Singhal, M., 2021) отмечают, что инициализация пустых кластеров обнаруживается в результате применения алгоритма случайного леса. Разработке инструмента визуализации точечных диаграмм рассеивания с выделением пустых кластеров посвящена статья (Giesen J. и др., 2015)

Для учета эффекта наличия пустых кластеров предлагается процедура внедрения (Audigier и др., 2021). Его идея состоит в том, чтобы сначала заполнить набор данных наблюдениями, соответствующими в противном случае неидентифицируемым кластерам. Недостаток подхода заключается в том, что он требует предварительного знания того, куда добавляются новые наблюдения. Проблеме пустых кластеров, также посвящены две статьи (Forina M. и др., 2003) в которых авторы предлагают статистический тест для оценки качества агломеративной кластеризации, а также предлагают индекс информативности «пустых пространств», учитывающий наибольшую дистанцию между кластерами. (McGee G. и др., 2020) исследуют долгосрочные эффекты на здоровье, в частности, на рождаемость детей, в результате внешних воздействий и отмечают, что, подходы с использованием оценочных уравнений обязательно исключают пустые кластеры и, следовательно, дают предвзятые оценки предельных эффектов.

1. **Методология научной работы**

Для проверки гипотезы о работоспособности алгоритма использовался тестовый набор данных из 30 наблюдений, заданный вручную, для явного выделение пропуска, или же, в нашей нотации, пустого кластера. Далее в этот набор был добавлен равномерно распределенный шум. Для последующей апробации использовался генератор данных *make\_blobs* библиотеки *sklearn*. Для удобства все сгенерированные данные были приведены к целочисленному типу. В одномерном варианте такие наборы данных можно интерпретировать как смеси нормальных распределений с заданной модой и стандартным отклонением. Так как наша задача предполагает вручную заданное число кластеров, то для поиска оптимального изначального разделения мы использовались следующие методы: Distortion Measure (Elbow-method), Silhouette Analysis, Calinski Harabasz, Gap-statistics.

В работе предлагается рассмотреть трехшаговую процедуру по инициализации и заполнению пустого кластера и оценки влияния его наличия. Предполагается, что те наблюдения, которые при новом разбиении на кластеры попали в отличный от первого кластер должны подлежать дополнительной проверке. Процедура состоит из следуюших этапов:

* на первом шаге предлагается оценить диапазон количества кластеров;
* на втором шаге, исходя для всего диапазона количества кластеров с предыдущего шага и их характеристик (минимальное, максимальное, среднее значение внутри кластера, количество наблюдений в кластере), оценивается размер пустого кластера и его характеристики – предполагаемый центорид, количество наблюдений и разброс значений. Генерируется набор данных с характеристиками пустого кластера;
* на третьем шаге проводится повторная инициализация кластеризации с добавлением «пустого» кластера в исходный массив и сравниваются результаты кластеризации с первым. Выделяются те наблюдения, чья метка кластера отличается от первой процедуры кластеризаци.

1. **Основные полученные результаты работы и новизна**

Реализован алгоритм выделения пустого кластера и его генерации исходя из предполагаемых характеристик. При варьировании размера пустого кластера при постоянных моде и стандартном отклонении отмечается, что если: его размер и ширина меньше среднего у исходных, то центроиды практически не смещаются относительно исходных, если равно или больше – то наоборот, наблюдается смещение центроидов и больше объектов получает новую метку. Таким образом алгоритм выделения и генерации пустого кластера должен быть параметрическим.

В дальнейшем исследовании предлагается:

* Добавлять пустые кластеры не только внутри существующего пространства, но и за границами исходного разделения;
* Разработать алгоритм для многомерной кластеризации с разными метриками расстояния между кластерами и их границами.

1. **Список литературы**

1. Feature Engineering v2.0 - Clustering With K-Means [Электронный ресурс]. URL: <https://kaggle.com/code/georgezoto/feature-engineering-v2-0-clustering-with-k-means> (дата обращения: 24.11.2022).

2. Xu D., Tian Y. A Comprehensive Survey of Clustering Algorithms // Ann. Data. Sci. 2015. Т. 2. № 2. С. 165–193.

3. Raschka, S. and Mirjalili, V. Python Machine Learning. Machine Learning and Deep Learning with Python, scikitlearn, and TensorFlow 2. Packt, 3rd edition, 2019

4. Yadav A., Dhingra S. An Enhanced K-Means Clustering Algorithm to Remove Empty Clusters // 2016. Т. 4. № 4. С. 7.

5. Pakhira M. A Modified k-means Algorithm to Avoid Empty Clusters // International Journal of Recent Trends in Engineering. 2009. Т. 1.

6. Raykov Y. P. и др. What to Do When K-Means Clustering Fails: A Simple yet Principled Alternative Algorithm // PLOS ONE. 2016. Т. 11. № 9. С. e0162259.

7. Hua C. и др. A Genetic XK-Means Algorithm with Empty Cluster Reassignment // Symmetry. 2019. Т. 11. № 6. С. 744.

8. Tavallali, P., Singhal, M. K-means tree: an optimal clustering tree for unsupervised learning. The Journal of Supercomputing, 77:5239–5266, 2021.

9. Giesen J., Kühne L., Lucas P. Sclow Plots: Visualizing Empty Space // Computer Graphics Forum. 2017. Т. 36. № 3. С. 145–155.

10. Audigier V., Niang N., Resche-Rigon M. Clustering with missing data: which imputation model for which cluster analysis method? // 2021.

11. Forina M., Casolino C., Lanteri S. Cluster analysis: significance, empty space, clustering tendency, non-uniformity. I--Statistical tests on the significance of clusters // Ann Chim. 2003. Т. 93. № 1–2. С. 55–68.

12. Forina M., Lanteri S., Casolino C. Cluster analysis: significance, empty space, clustering tendency, non-uniformity. II--Empty Space index // Ann Chim. 2003. Т. 93. № 5–6. С. 489–498.

13. McGee G. и др. Informatively empty clusters with application to multigenerational studies // Biostatistics. 2020. Т. 21. № 4. С. 775–789.

14. Marutho D. и др. The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News // 2018 International Seminar on Application for Technology of Information and Communication. , 2018. С. 533–538.

15. Shutaywi M., Kachouie N. N. Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering // Entropy. 2021. Т. 23. № 6. С. 759.

16. Wang X., Xu Y. An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index // IOP Conf. Ser.: Mater. Sci. Eng. 2019. Т. 569. № 5. С. 052024.

17. Estimating the number of clusters in a data set via the gap statistic - Tibshirani - 2001 - Journal of the Royal Statistical Society: Series B (Statistical Methodology) - Wiley Online Library [Электронный ресурс]. URL: <https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00293> (дата обращения: 19.11.2022).