Метод k-means.

При реализации метода сталкиваюсь с несовпадением разбиения, чукотку всегда выделяет в 1 кластер, а Москву закидывает к Магадану.

При установке рандомного выбора центроидов - для 2020, 2019 получаем эталонное разбиение

Рассмотрим аргументы kmeans:

**Init - отвечает за то, как будут выбраны центроиды**

1. “k-means++” - такой параметр инициализирует центроиды удалёнными друг от друга, что помогает найти глобальный минимум (алгоритм сходится к нему) - такой параметр не подходит для моей задачи, потому что три кластера лежат довольно близко друг к другу, к тому же есть субъект, который сильно отдалён от других и, в данном случае, будет включён в отдельный кластер. (Чукотский АО)
2. “random” - центроиды будут выбираться рандомно, что может грозить неправильной кластеризацией.

PCA (анализ главных компонент) поможет сократить кол-во признаков до двух (1 - экономический, 2 - инновационный).

**n\_init** - сколько раз будут выбираться центроиды (выбирается лучший вариант с точки зрения инерции)

* Помогает предотвратить проблему того, что элемент может оказаться изолированным от других и образовать кластер из одного элемента.

Список литературы

12. A. K. Jain and R. C. Dubes, Algorithms for Clustering Data, Englewood Cliffs, NJ, USA:Prentice-Hall, 1988 - книга, которая часто используется в статьях

Data Clustering: A Review A.K. JAIN

G. N. Lance and W. T. Williams, “A General Theory of Classificatory Sorting Strategies 1. Hierarchical Systems,” The Computer Journal, Vol. 9, No. 4, 1967, pp. 373-380.  
<http://dx.doi.org/10.1093/comjnl/9.4.373>

Литература на Штейнгауза и Ллойда про K-Means

13. Steinhaus H. Sur la division des corps materiels en parties // Bull. Acad. Polon. Sci. C1. III. 1956. Vol. IV. P. 801–804.

14. Lloyd S. Least squares quantization in PCM // IEEE Transactions on Information Theory. 1982. Vol. 28. Iss. 2. P. 129–137. doi: 10.1109/TIT.1982.1056489

Про приемущества иерархического метода одиночной связи:

16. Aldenderfer M. S., Blashfield R. K. Cluster analysis. Newburg Park: Sage Publications Inc., 1984. 88 p

17. Орехов А. В. Марковский момент остановки агломеративного процесса кластеризации в евклидовом пространстве // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2019. Т. 15. Вып. 1. С. 76–92. <https://doi.org/10.21638/11702/spbu10.2019.106>

18. Виолин Сергей Игоревич

Типологизация регионов как основа для проведения дифференцированной государственной региональной политики// Региональная экономика и управление: электронный научный журнал. ISSN 1999-2645. — №2 (54). Номер статьи: 5406. Дата публикации: 18.04.2018. Режим доступа: <https://eee-region.ru/article/5406/>

19. Орлова И.В., Филонова Е.С. КЛАСТЕРНЫЙ АНАЛИЗ РЕГИОНОВ ЦЕНТРАЛЬНОГО ФЕДЕРАЛЬНОГО ОКРУГА ПО СОЦИАЛЬНО-ЭКОНОМИЧЕСКИМ И ДЕМОГРАФИЧЕСКИМ ПОКАЗАТЕЛЯМ. Статистика и Экономика. 2015;(5):111-115. <https://doi.org/10.21686/2500-3925-2015-5-136-142>

20. Алтунина В.В., Анучина Д.А. Классификация регионов Российской Федерации в контексте пространственной поляризации // Экономика, предпринимательство и право. – 2022. – Том 12. – № 5. – С. 1453-1474. – doi: 10.18334/epp.12.5.114641.

21. Шматко А.Д., Губин С.В. Кластерный анализ инновационного потенциала субъектов РФ. Управленческое консультирование. 2020;(3):61-72. https://doi.org/10.22394/1726-1139-2020-3-61-72

22. Шамрай-Курбатова Л. В., Леденева М. В. Кластерный анализ субъектов РФ по уровню инновационной активности // Бизнес. Образование. Право. 2021. № 1 (54). С. 88—97. DOI: 10.25683/VOLBI.2021.54.174.

23. Исследование инновационного развития и типология регионов на основе интегральной оценки их конкурентоспособности // Известия Уральского государственного экономического университета. 2011. № 2. С. 83–91.

24. Мыслякова Ю.Г. Разработка типологии регионов по их предрасположенности к научно-технологическому развитию. Экономика и управление. 2021;27(10):775-785. <https://doi.org/10.35854/1998-1627-2021-10-775-785>

25. Заварухин

26. Рейтинг инновационного развития субъектов Российской Федерации. Выпуск 8 / В. Л. Абашкин, Г. И. Абдрахманова, С. В. Бредихин и др.; под ред. Л. М. Гохберга; Нац. исслед. ун-т «Высшая школа ­экономики». – М. : ИСИЭЗ ВШЭ, 2023. – 260 с. – 80 экз. – ISBN 978-5-7598-3000-9 (в обл.)

27. D. Arthur, S. Vassilvitskii

K-means++: the advantages of careful seeding

Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms SODA ’07, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA (2007), pp. 1027-1035

28. Старовойтов, В. В. Нормализация данных в машинном обучении / В. В. Старовойтов, Ю. И. Голуб // Информатика. – 2021. − Т. 18, № 3. – С. 83–96. https://doi.org/10.37661/1816-0301-2021-18-3- 83-96

План.

1. Определение задачи кластеризации. Здесь рассмотреть типы и вывести хорошее определение. Тут же подойти к рассмотрению разных методов кластерного анализа. (3-4 станицы)
2. Потом начать описание методов каждого по отдельности с формулами через матрицу сходства и т д (7 –8 страниц)
3. Описать метрики для определения правильности разбиения, описать что разбиение буду осуществлять несколько раз для определения наиболее часто встречающегося. (3 страницы)
4. Затем описание применения каждого из методов к моим данным + проверка по метрикам. (5 страниц) тут возникнет вопрос сколько данных мне нужно для правильного разбиения (верное разбиение уже есть)

2 раздел - Метод К - средних, определение числа кластеров, реализовать метод Орехова и еще один, та же тема с объемом данных.

В этом разделе необходимо программно реализовать методы нахождения числа кластеров и провести машинные эксперименты.

3 раздел - Кластеризация матриц, каждый столбец - отдельный год, здесь Вы можете придумать свой алгоритм кластеризации, может быть, не один.