САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Направление: 01.03.02 «Прикладная математика и информатика»
ООП: Прикладная математика, фундаментальная информатика и программирование

ОТЧЕТ О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ ПРАКТИКЕ

Тема задания: Сравнение различных методов кластеризации регионов по экономическим и инновационным показателям.

Выполнил: Дашкова О.В. 20.Б02-пу

Руководитель практики
от СПбГУ: доктор технических наук, Буре В.М.

Санкт-Петербург
2023

Введение.

Задача кластеризации является одной из наиболее популярных в различных сферах исследования. Существует множество различных методов и алгоритмов. Некоторые из них могут подойти для решения одной поставленной задачи, но при этом быть непригодными в использовании для иного разбиения. Поэтому рассмотрение и сравнение каждого из них довольно таки трудоемкий и кропотливый процесс. Более того, для качественного анализа методов кластеризации необходим пригодный для использования набор данных. Если обратиться к определению - задача кластерного анализа заключается в том, чтобы на основании данных, каких - либо наблюдаемых показателей или характеристик, разбить множество объектов на k кластеров (подмножеств) так, чтобы каждый объект принадлежал одному и только одному подмножеству разбиения и чтобы объекты принадлежащие одному и тому же кластеру, были сходными, в то время как объекты, принадлежащие разным кластерам были разнородными.[1] Условно методы кластеризации можно разбить на две группы - иерархические и неиерархические. Первые заключаются в построении так называемого дерева вложенных кластеров, основываясь на различных способах вычисления расстояния между ними. Вторые же более удобны в случае большого количества объектов, подлежащих объединению, и строят одно разбиение на кластеры. В качестве исходных данных выбрана официальная информация Росстата по 87-ми субъектам РФ по следующим экономическим и инновационным показателям за 2019 и 2020 года: ВРП на душу населения, инвестиции в основной капитал на душу населения, объем основных фондов на душу населения, коэффициент изобретательной активности, внутренние затраты на исследования и разработки в расчете на одного работника, занятого научными исследованиями, объем отгруженных инновационных товаров работ и услуг на душу населения. Информационной базой для такого выбора послужила рассматриваемая мною статья [6].
Так как некоторые данные по инновационным признакам не публикуются в целях обеспечения конфиденциальности, были рассмотрены 82 субъекта за 2019 год и 77 субъектов за 2020 год. Так же один субъект был убран, так как вызывал сильные выбросы, что препятствовало правильной работе метода k-means. По итогу построение всех методов было осуществлено на основе экономико-инновационных показателей для 77 субъектов РФ.

Задачи.

Целью исследования являлось применение различных методов кластеризации, их сравнение и выявление наиболее подходящего из них для рассматриваемой задачи разбиения субъектов по экономическим и инновационным показателям. Также, после создания кластерного решения, требуется оценка его качества. Данная задача трудновыполнима, так как не существует оптимального алгоритма кластеризации.

Модели, методы, алгоритмы.

В качестве неиерархических методов рассмотрены: метод k-means, k-medoids, алгоритм CLARA. В качестве иерархических: агломеративные (расстояние между кластерами: средняя связь, метод ближнего соседа, метод дальнего соседа) и дивизионные методы (DIANA). При использовании неиерархических методов, необходимо знать число кластеров. Для этого используется несколько алгоритмов и более того, сама задача отыскания оптимального числа кластеров является актуальной и требует отдельного рассмотрения. Но мы остановимся на наиболее популярных индексных методах: статистика разрыва (gap statistic method), метод локтя (elbow method), метод силуэтов (silhouette method). Реализация всех методов осуществлялась на языке программирования R; визуализация реализована на основе графической системы ggplot2 при помощи пакета "factoextra". В качестве методов для оценки качества кластеризации могут быть использованы внешняя, внутренняя валидация и оценка стабильности объединения в кластеры[7].

Предварительные результаты.

Были рассмотрены и реализованы основные типы методов, на основании которых можно проводить дальнейший анализ. Иерархические алгоритмы сравнили между собой, используя коэффициент W матричной корреляции Мантеля. Пришли к выводу, что наибольшая адекватность кластерного решения исходным данным принадлежит кластеризации по методу средней связи (рассчитывает расстояние между кластерами как среднее арифметическое между всеми парами объектов). Сравнение же остальных типов кластеризации не было проведено, но может быть осуществлено в дальнейших исследованиях.

Литература.

1. B.S. Duran, P.L. Odell "Cluster Analysis: A Survey 1974. Перевод с английского: Е.З. Демиденко, с. 15;
2. И.Д. Мандель "Кластерный анализ 1988
3. В.М. Буре, Е.М. Парилина "Теория вероятностей и математическая статистика"
4. В.М. Буре, Е.М. Парилина, А.А. Седаков "Методы прикладной статистики в R и Excel"
5. «Кластерный анализ: Википедия. Свободная энциклопедия.
Режим доступа: https://ru.wikipedia.org/wiki/Кластерный анализ (дата обращения: 31.05.2023)»
6. В.П. Заварухин, Т.И. Чинаева, Э.Ю.Чурилова "Регионы России: результаты кластеризации на основе экономических и инновационных показателей"// Статистика и экономика Т.19, №5, 2022 7. Оценка качества кластееризации. Режим доступа: https://ranalytics.github.io/data-mining/103-Clustering-Quality.html