Верефикация кластеризации временных рядов

Кривонос Анна

Московский государственный университет им. М. В. Ломоносова

Курс: Моя первая научная статья

Научный руководитель: О. В. Сенько

2023

Введение

Важнейшим задачами эпидемиологии являются исследования влияния различных факторов на ход эпидемического процесса, а также прогнозирование развития эпидемии. Например, заболеваемость коронавирусной инфекцией (COVID-19), поразившей практически весь мир в 2019-2021 годах, протекала в разных регионах и странам мира по-разному в зависимости от состояния систем здравоохранения, климатических, социально-экономических, демографических условий, других характеристик регионов. Для решения обоих задач могут быть применены современные методы машинного обучения и анализа данных. Целью настоящей работы является поиск оптимальной схемы использования кластерного анализа, являющегося популярным и эффективным инструментов современного анализа данных, для изучения эпидемического процесса.

Постановка задачи

В качестве данных в даннай задече используются кривые темпа роста Covid-19 для различных странн мира, а также для отдельных регионов России.
Необходимо провести кластеризацию временных рядов, которая позволяла бы оценить ход эпидемиологического процесса и влияние на него различных факторов. Для этого проводится верефикация полученной кластеризаци. Для верификации кластеризации рассматривается подход, основанный на проверке нулевой гипотезы о равновероятности различных соответствий мер сходства между двумя временными рядами.

Описание метода кластеризации

На первом этапе вычислялась мера сходства между всевозможными парами эпид-кривых S_i и S_j через подбор лага l из отрезка [0,20].
Для каждого l вычислялся коэффициент корреляции p_l^+ между рядами $S_i(0), ..., S_i(n-l)$ и $S_j(l), ..., S_j(n)$ и коэффициент корреляции p_l^- между рядами $S_i(l), ..., S_i(n)$ и $S_j(0), ..., S_j(n-l)$.
$p(S_i, S_j) = max\{p_0^+, p_0^-, ..., p_{20}^+, p_{20}^-\}$
Пусть P_{m*m} является матрицей сходства m эпид-кривых. На диагонали симметричной матрицы P_{m*m} находятся единицы, а внедиагональными элементами являются максимальные коэффициенты корреляции $p(S_i, S_j)$, рассчитанные согласно приведённой выше процедуре.

Описание метода кластеризации

После подсчёта мер близости между кривыми использовался метод иерархической агломеративной кластеризации. В качестве меры сходства двух групп эпидкривых $G^{'}$ и $G^{''}$ использовалось среднее значение меры сходства между эпидкривыми из разных групп:

$$P(G^{'}, G^{''}) = \frac{1}{m^{'} m^{''}} \sum_{i=1}^{m^{'}} \sum_{j=1}^{m^{''}} p(S_i, S_j)$$

Процесс слияния кластеров прекращался, если мера сходства P между любыми двумя кластерами в текущей кластеризацией не окажется ниже 0.5.

Метод верификации

Пронумеруем элементы матрицы P_{m*m}, находящиеся выше диагонали.
Пусть I взаимно однозначное отображение множества $\{(i,j)|i,j=1,...,m,i<j\}$ в $\{1,...,M\}$, где $M=\frac{m(m-1)}{2}$
Пусть f – перестановка элементов множества $\{1,...,M\}$.
По перестановке f строится матрица P^f_{m*m}. Пусть $k=I(i,j)$.
Тогда элементу матрицы P^f_{m*m} в позиции (i,j) присваивается элемент P_{m*m}, который находится в позиции $(i^*,j^*)=I^{-1}[f(k)]$.
Элементы ниже главной диагонали запоняются симметричным образом.

Метод верификации

Значения индикатора качества $I(C)$ для кластеризации C, полученной по исходной матрице сходства P_{m*m} сравнивается со значениями индикатора качества $I(C^f)$ для кластеризаций C^f, по матрицам сходства P^f_{m*m}, генерируемым по $\widetilde{f}$ - случайному подмножеству переестановок элементов множества $\{1, ..., M\}$.

В качестве $\rho-$значения ипользуется доля перестановок, при которых значение индикатора качества $I(C^f)$ для кластеризации C^f достигает или превосходит значения индикатора качества $I(C)$ для кластеризации C,

$$\rho = \frac{|\left\{f \in \widetilde{f} | I(C^f) \geq I(C)\right\}|}{|\widetilde{f}|}$$

Метод верификации

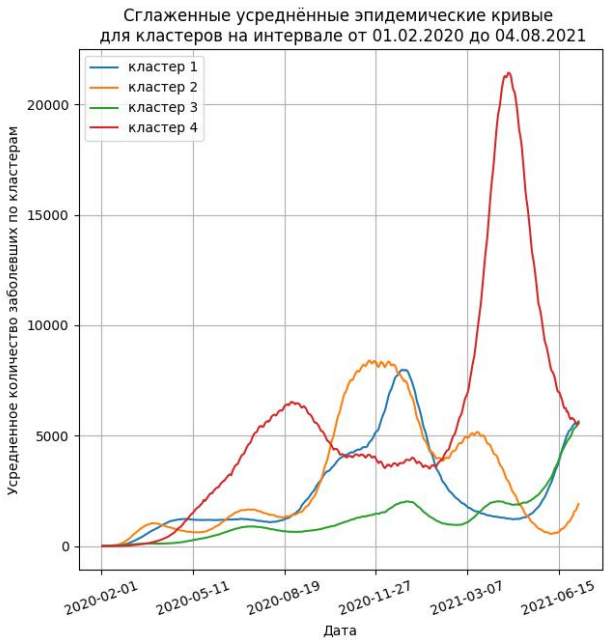
В качестве индикатора качества предлагается воспользоваться коэффициентом силуэта.

$$I(n) = \frac{b(n) - a(n)}{max\{a(n), b(n)\}},$$

где $a(n)$ - среднее расстояние от n до другого объекта внутри кластера, $b(n)$ - среднее расстояние от n до объекта в другом кластере. Среднее значение всех силуэтов называют коэффициентом силуэта. В нашем случае в качестве метрики для оценки силуэта используется максимальный коэффициент корреляции Пирсона с лагом:

$$d_{i,j} = \rho_{max}(S_i, S_j)$$

Эксперименты



- кластер 1: Великобритания, Россия, США, Португалия и др.
- кластер 2: Австрия, Германия, Италия, Сербия, Болгария, Румыния и др. Географически локализованы главным образом в Европе.
- кластер 3: Алжир, Марокко, Бангладеш, Вьетнам и др. Страны с тропическим климатом.
- кластер 4: Аргентина, Колумбия, Парагвай, Уругвай, Камбоджа и др. Географически локализована в Южной Америки.

Эксперименты

Для данных о динамике заболеваемости за временной промежуток от 22-01-2020 до 05-08-2021 среднее значение всех вычисленных силуэтов при разделении стран на 4 обозначенных кластера составило 0.36.

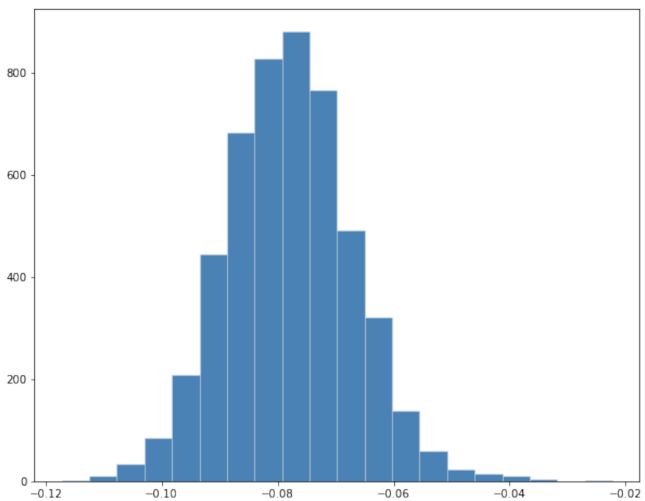


Рис. 5: Коэффициенты силуэта для 5000 различных экспериментов.

Для верификации результата предлагается вычислить коэффициент с помощью использования перестановок. Осуществим 5000 различных перестановок f и посчитаем коэффициенты силуэта. Максимальное значение коэффициента силуэта равно 0.02.

Эксперименты

Проведем кластеризацию для случайных сгенерированных временных рядов из нормального распределения. Полученный коэффициент силуэта на основе алгоритма 3 для нашей кластеризации составил -0.004.

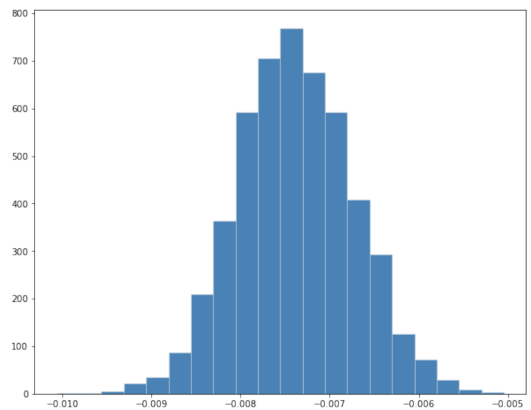


Рис. 6: Коэффициенты силуэта для 5000 различных экспериментов для случайных сгенерированных рядов.

Не смотря на отрицательное значение коэффициента силуэта он все же остается выше, чем коэффициенты для полученных перестановок - максимальный коэффициент силуэта для 5000 перестановок составил -0.005.

Выводы

В работе описан метод кластеризации временных рядов, приведен способ оценки значимости кластеризации. В ходе экспериментов показана хорошая работа на эпидемиологических кривых Covid-19. В дальнейшем планируется усовершенствовать метод кластеризации с помощью введения остановки слияния кластеров в заыисимости от вычисляемых ρ-значений для каждого кластера на каждом этапе слияния.