Применение активного обучения к графовым моделям на примере оценки рисков распространения эпидемии

Бишук Антон Юрьевич

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель к.ф.-м.н. Зухба А.В.

Москва
2021 г

Цели и задачи

Цели

- Оценить влияние ограничительных мер.
  1. Для монотонных – предложить оптимальную стратегию.
  2. Для немонотонных – предложить критерий положительного влияния.
- Предложить стратегию, позволяющую оценивать эпидемиологическую обстановку без массового тестирования.

Задачи

- Поиск условий, при которых локдаун может увеличивать математическое ожидание числа заболевших;
- Поиск стратегии вакцинации, минимизирующей число заболевших;
- Поиск стратегии тестирования, оптимизирующей информационные критерии системы при условии ограниченного количества тестов.

Граф контактов

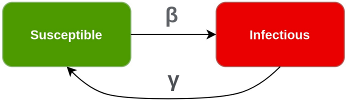
Граф контактов – граф, у которого вершина – человек, а ребро – контакт между людьми. Ребра могут иметь веса: продолжительность контакта и т.д.

- Каждая вершина графа может находиться в конечном количестве состояний: болен, здоров, уязвим, и т.д
- Состояния вершин, а также веса ребер меняются во времени.
- Состояние системы определяется состоянием всех вершин.

Итерация (шаг моделирования) – в данной работе временной шкалой принято считать дни, а значит за один шаг (одну итерацию) принимается один день.

Будем называть ***величиной окрестности степени*** k ***вершины*** v – число вершин, до которых можно добраться следуя по ребрам исходного графа не более чем за k переходов.

Модель SIS

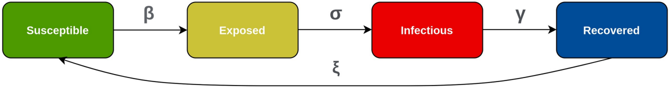


$$\begin{cases} \frac{dS}{dt} = \frac{-\beta SI}{N} + \gamma I, \\ \frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I. \end{cases}$$

Где $S + I = N$ – вся популяция

β – вероятность заразиться,
γ – вероятность выздороветь.

Модель SEIRS



$$
\begin{cases}
\frac{dS}{dt} = -\frac{\beta IS}{N} + \xi R, \\
\frac{dE}{dt} = \frac{\beta IS}{N} - \sigma E, \\
\frac{dI}{dt} = \sigma E - \gamma I, \\
\frac{dR}{dt} = \gamma I - \xi R.
\end{cases}
$$

Где $S + E + I + R = N$ – вся популяция

β – вероятность заразиться,
σ – вероятность перейти из состояния зараженного в состояние больного,
γ – вероятность выздороветь и приобрести иммунитет,
ξ – вероятность потерять иммунитет.

Теорема об «эквивалентности» вершин

Theorem (Bishuk, 2021)

Пусть дан граф $G(V, E)$, пусть в нем $n > 0$ вершин заражено, тогда число заболевших на следующей итерации монотонно от величины окрестности для всех больных вершин.

$$\mathbb{E}(I_1) = \sum_{i=1}^{k} (i-1) \cdot \sum_{j=1}^{C^i_{k+1}} p_{i,j}$$

Motivation

При моделировании, случайная инициализация больных приводит к слишком большому разбросу в поведении эпидемии. Представленная теорема обосновывает использование окрестностей, а в первом приближении степеней, вершин для стратификации.

Противоэпидемиологические меры для графовых структур

Для борьбы с распространением эпидемии используется ряд противоэпидемиологических мер:

- Вакцинация (Смена статуса вершины $S \to R$),
- Изоляция больных (Удаление вершины и инцидентных с ней рёбер),
- Введение lockdown (Смена графа контактов на граф, состоящий из клик небольшого размера).

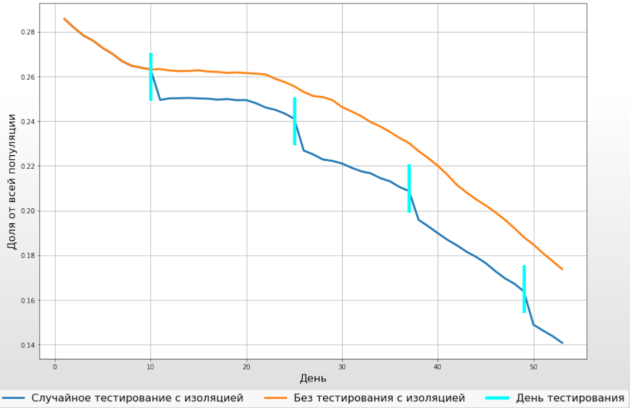
И если использование первых двух способов сдерживания эпидемии в худшем случае могут не возыметь эффект, то введение локдауна при определенных условиях может вызвать еще больший рост эпидемии.

- Тестирование на течение эпидемии непосредственно не влияет, но является необходимым инструментом для оценки ситуации.

Монотонные эпидемиологические меры: тестирование с изоляцией

Lemma

Тестирование с последующей изоляцией больных не увеличивает число заболевших на следующей итерации.



Монотонные эпидемиологические меры: вакцинация

Lemma

Вакцинация не увеличивает число заболевших на следующей итерации.

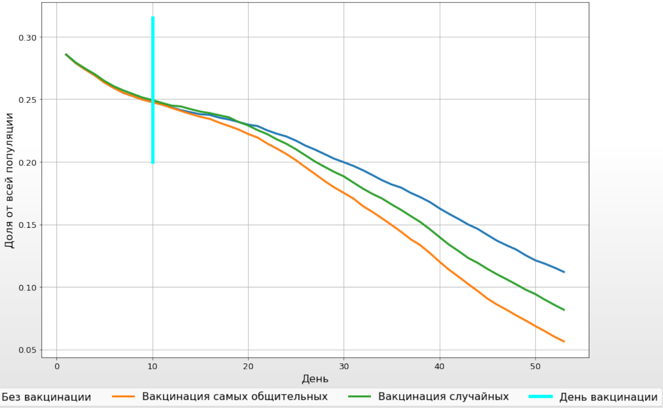


Рис.: Среди всех стратегий вакцинации наиболее эффективной оказалась вакцинация самых общительных – с наибольшим числом контактов в прошлом.

Обозначения

$G^H(V,E)$, $G^W(V,E)$ – графы контактов дома и работы соответственно.

$I_t(A)$ – подмножества вершин из множества вершин $A \subseteq V$, находящиеся в момент времени t в состояниях Infectious.

$V_i = \{v_j | e_{ij} > 0\}$ – множество контактных с v_i вершин.

Вероятность заразиться на следующей итерации:

$$\beta_t^i = \mathbb{P}(v_i \in E_{t+1} | v_i \in S_t) = |I_t(V_i)| \frac{1}{|V_i|} B_i$$

Пусть β линейно зависит от времени проведенного в графе дома и работы соответственно через B_i^H и B_i^W:

$$\beta_t^i = |I_t(V_i)| \frac{1}{|V_i|} \tilde{B}_i \cdot \tau$$

Увеличение вероятности заразиться при локдауне

Lemma (Bishuk, 2021)

В случае линейной зависимости заражения от времени, шанс заразиться для вершин v_i растет при локдауне тогда и только тогда, когда

$$\frac{|I_k(V_i^H)|\tilde{B}_i^H}{|V_i^H|} \geq \frac{|I_k(V_i^W)|\tilde{B}_i^W}{|V_i^W|}$$

Corollary

Если есть только информация о том, какой процент населения болен (обозначим это значение как q), то можно выделить следующий вид условия леммы:

$$\frac{\tilde{B}_i^H}{\tilde{B}_i^W} \geq \frac{1-(1-q)^{|V_i^H|}}{1-(1-q)^{|V_i^W|}}$$

Критерий «плохого» локдауна

Theorem (Bishuk, 2021)

Локдаун уменьшает математическое ожидание числа заболевших тогда и только тогда, когда:

$$\mathbb{E}(\tilde{I}_{k+1} - I_{k+1}) = \sum_{i: v_i \in I_k} (\tilde{\beta}_i^k - \beta_i^k) =$$

$$= \sum_{i=1}^{|V|} [v_i \in S_k] \frac{B_i^H |V_i^W| |I_k(V_i^H)| - B_i^W |V_i^H| |I_k(V_i^W)|}{|V_i^H| \cdot |V_i^W|} < 0$$

Доказательство.

Является прямым следствием описанный ранее леммы. □

Локдаун при правильной стратегии введения

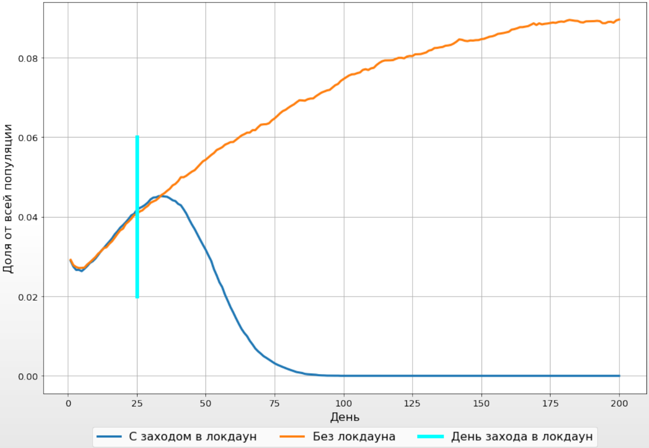


Рис.: После введения локдауна число заболевших начало резко падать, то есть, локдаун помог в сдерживании распространении эпидемии.

Локдаун при неправильной стратегии введения

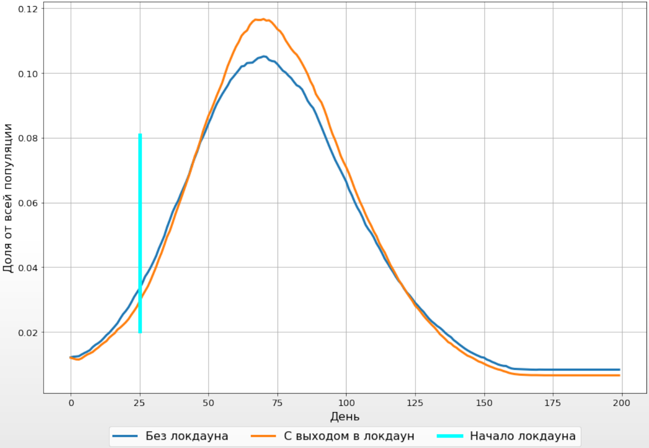


Рис.: После введения локдауна число заболевших начало постепенно расти и в пике, их было на $\sim 1.5\%$ больше, чем в ситуации когда локдаун не был введен.

Модель стратегий тестирования

- Рассматриваем модель SIS.
- Функция пересчета вероятностей (рисков):

$$p_i^t = p_i^{t-1}(1-\gamma) + (1-p_i^{t-1}) f(\beta, \{p_{i_k}\})$$

- Задача выбора людей для тестирования $\rightarrow$ задача ранжирования по приоритету.
- Для ранжирования могут использоваться различные критерий. В частности, критерии, основанные на энтропии.

Энтропия и пересчет вероятностей

Граф это система, в вершинах, которой стоят вероятности того, что она больна (далее «вероятность вершины»).

Тогда под энтропией графа будем понимать следующую величину:

$$\mathbb{H}(\{x_i\}_{i=1}^n) = -\left(\sum_{i=1}^n p_i \log(p_i) + \sum_{i=1}^n (1-p_i)\log(1-p_i)\right) = \sum_{i=1}^n H_i$$

Если известно распределение вероятностей на прошлой итерации $\{p_i^{t-1}\}_{i=1}^n$, то пересчёт вероятностей на следующей итерации:

$$p_i^t = p_i^{t-1}(1-\gamma) + (1-p_i^{t-1})\left(\frac{2\arctan\left(\sum_k p_{i_k}\right)^c}{\pi}\right)^{-\log_2(\beta)} \tag{1}$$

где p_{i_k} – вероятности соседей вершины v_i

Алгоритм поиска вершины с большим вкладом в энтропию $(O(E+V))$

1. Каждая вершина помнит своих соседей и их вероятности с прошлого шага.
2. Пересчитывается вероятности согласно формуле (1).
3. Для каждой вершины на основе новых вероятностей считается H_i и величина $H_i^t + \sum_{i_k} H_{i_k}^t$.
4. Для каждой вершины вычисляется:
$$p_i \sum_{i_k} \tilde{H}_{i_k} + (1-p_i) \sum_{i_k} \tilde{\tilde{H}}_{i_k}.$$
5. И тогда необходимо найти ту вершину, для которой величина
$$\left| H_i^t + \sum_k H_{i_k}^t - (p_i \sum_{i_k} \tilde{H}_{i_k} + (1-p_i) \sum_k \tilde{\tilde{H}}_{i_k}) \right|$$
будет максимальной.

Такой критерий поиска будем называть ***критерий изменения энтропии.***

NDCG

$$NDCG@k = \frac{\sum_{i}^{k} G_q(d_q^{(i)}) \cdot D(i)}{\max(\sum_{i}^{K} G_q(d_q^{(i)}) \cdot D(i))}, \text{ где } G_q(d) = 2^{y(q,d)} - 1, \quad D(i) = \frac{1}{\log_2(i+1)}$$

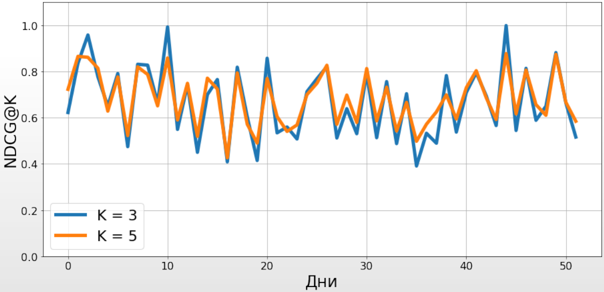
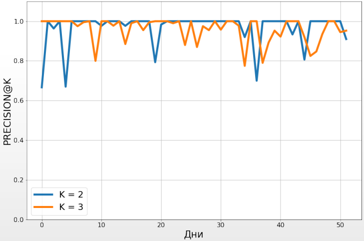


Рис.: Метрика точности ранжирования вершин по критерию изменения энтропии.

PRECISION

$$PRECISION@k = \frac{\text{count}(\text{Top-k predicted rank})}{\text{count}(\text{Top-k predicted rank} \cup \text{Top-K true rank})}$$



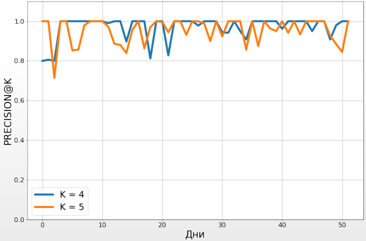


Рис.: Метрика точности ранжирования вершин по критерию изменения энтропии.

MAE и вероятностный критерий 1

Вероятностный критерий 1:

$$P_i = \left|p_i - \frac{1}{2}\right| + \sum_k \left|p_{i_k} - \frac{1}{2}\right| \to \max_i$$

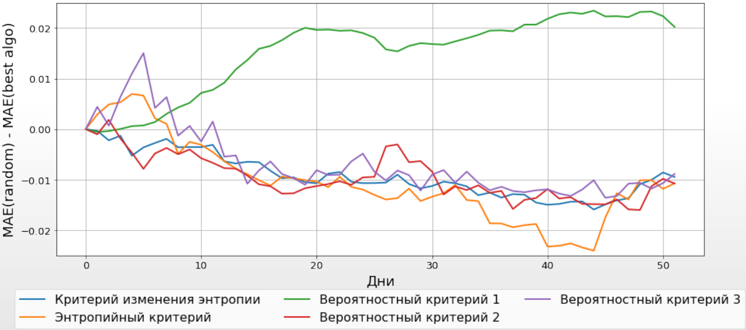


Рис.: random – тестирование случайного человека на каждом шаге, best algo – тестирование лучшего человека с точки зрения критерия. MAE(·) – суммарная абсолютная ошибка предсказания меток для всех вершин в конкретный день.

Выводы и результаты

- Изучен эффект отрицательного влияния локдауна. Сформулирован и доказан критерий увеличения математического ожидания числа заболевших при локдауне.
- Сформулирована и доказана теорема об эквивалентности людей для симуляции эпидемии.
- Сформулированы и экспериментально доказаны леммы о монотонных ограничительных мерах.
- Сформулирована задача в терминах теории информации.
- Предложен, реализован и экспериментально проверен алгоритм тестирования, основанный на нескольких информационных критериях.