Проблемы Формирования Репрезентативной Выборки для Выделения Значимых Признаков COVID-19

Юрий Мищеряков
кафедра системотехники
Харьковский Национальный Университет Радиоэлектроники
Харьков, Украина
iurii.mishcheriakov@nure.ua

Дмитрий Ситников
кафедра системотехники
Харьковский Национальный Университет Радиоэлектроники
Харьков, Украина
dmytro.sytnikov@nure.ua

Михаил Ищенко
кафедра системотехники
Харьковский Национальный Университет Радиоэлектроники
Харьков, Украина
mykhailo.ishchenko@nure.ua

Александр Украинец
кафедра системотехники
Харьковский Национальный Университет Радиоэлектроники
Харьков, Украина
oleksandr.ukrainets@nure.ua

Problems of Forming a Representative Sample to Highlight Significant Signs of COVID-19

Iurii Mishcheriakov
dep. of Systems Engineering
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
iurii.mishcheriakov@nure.ua

Dmytro Sytnikov
dep. of Systems Engineering
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
dmytro.sytnikov@nure.ua

Mykhailo Ishchenko
dep. of Systems Engineering
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
mykhailo.ishchenko@nure.ua

Oleksandr Ukrainets
dep. of Artificial Intelligence
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
oleksandr.ukrainets@nure.ua

***Аннотация*—В данной работе анализируется проблема формирования информативных и репрезентативных выборок, на основе которых можно моделировать распространение вируса Covid-19. Особое внимание уделяется качеству данных и техническим аспектам их обработки, а также специфике тестирования.**

***Abstract*—In this work the problem of forming informative and representative samples is analyzed. Such sample can be used for Covid-19 distribution modelling. Special attention is paid to data quality, technical. aspects of data processing, and peculiarities of testing.**

Ключевые слова—Covid-19, репрезентативная выборка, моделирование, тестирование

Keywords— Covid-19, representative sample, modelling, testing

I. Введение

Репрезентативная выборка должна включать в себя прогнозные переменые и дополнительные параметры, влияющие на них. Выбор прогнозных переменных и дополнительных параметров зависит от цели. Допустим, необходимо спрогнозировать заболеваемость (заражение) и смертность в условиях пандемии COVID-19 в

зависимости от других факторов. Тогда прогнозными переменными будут «случаи заболевания» и «смертность». В качестве влияющих переменных может выступать огромное число факторов, таких как пол, возраст, расовая или этническая принадлежность, погодные условия, введенные властями ограничения в соответствующем регионе и многие другие. Выбор таких факторов зависит не только от цели моделирования, но и наличия статистических данных.

II. Основная часть

Одним из основных критериев качества выборки является её достаточность для выявления тенденций. По пандемии COVID-19, во многих странах, данные представлены с марта 2020 года и в основном в агрегированном виде, что является небольшой выборкой. Одним из способов увеличения объемов выборки это привязка по территориальному признаку в рамках страны, когда данные собираются с указанием места их получения. При этом одим из важнейших факторов, является одинаковость интерпретации собираемых данных. Здесь примером может быть выявление случаев одними и теми же методами во всех источниках данных.

Международные источники данных, такие как “European Centre for Disease Prevention and Control” [1], “The Institute for Health Metrics and Evaluation (IHME)” [2], “Our World in Data” [3, 4], “Johns Hopkins University” [5] [6] и другие, основными свойствами выделяют «случаи заболеваний» и «смертность» в привязке к стране, а для некоторых стран с привязкой к округу. Сопутствующие характеристики, такие как, например, распределение заболеваемости и смертности по возрастным категориям – отсутствуют. Информацию о таких характеристиках можно попытаться получить из официальных открытых источников конкретной страны. Например, США ведёт статистику заболеваемости также с учетом возрастных категорий и этнических признаков, однако если данные по заболеваемости и смертности предоставляются с учётом георасположения, то распределение по возрастным категориям и этническим признакам предоставляется без какой либо географической привязки. Таким образом, выявить влияние на различные возрастные категории введенных в регионе ограничений не представляется возможным.

Что касается Украины, то основными и, по-видимому, единственными источниками данных, являются «Центр громадського здоров’я України» [7] предоставляет данные в формате Excel и сайт Кабинета Министров Украины [8], предоставляет данные в формат csv. Данные на них совпадают. Представлено распределение с учётом разных видов тестирования (ПЦР, ИФА), а также распределение по типам реакции (IgA, IgM, IgG). При этом данные распределения по возрастным категориям в источнике не представлены, однако на дашборде представлены только агрегированные данные за весь период в процентном соотношении. Как и в предыдущем случае провести вышеуказанный анализ невозможно.

A. Качество представления исходных данных

Большое значение имеет человеческий фактор. Один из аспектов — это ввод текстовых характеристик, в частности, название расового / этнического признака. По данному признаку ярким примером может служить центр CDC, как один из основных источников данных США. С течением времени название одного и того же значения можен изменяться. Если рассмотреть данные в формате csv, предоставляемые Украиной, то названия лабораторий и названия столбцов могут содержать недопустимые символы, такие как перевод строки. Это всё требует дополнительной предобработки и тщательного контроля, перед внесением в БД.

B. Технические аспекты обработки исходных данных

Здесь существует множество факторов, затрудняющих внесение и дальнеёшую обработку данных. К ним следует отнести: различные форматы представления дат и чисел, человеческий фактор, внесение единиц измерения непосредственно в значение. Часто встречаются случаи, когда категориальные характеристики представлены как строки, а не как столбцы, т.е. для каждой даты будет содержаться несколько строк, каждая из которых соответстует своему категориальному значению. Иногда это выгодно, когда необходимо провести фильтрацию, но чаще всего, необходимо проводить транспонирование таких данных.

FIPS Code	Indicator	Total deaths	COVID-19 Deaths	Non-Hispanic White
1003	Distribution of all-cause deaths (%)	2009	104	0.897
1003	Distribution of COVID-19 deaths (%)	2009	104	0.808
1003	Distribution of population (%)	2009	104	0.832

Рис. 5. Категориальные характеристики в строках

C. Тесты

В данный момент в мире существует большое количество способов обнаружения коронавируса или остатков его деятельности в организме человека. Каждый тест имеет свои недостатки, будь то экономические или точностные, на которые предельно важно обратить внимание. У тестов зачастую имеется два параметра, отвечающих за точность определения, болен ли человек, и точность определения здоров ли человек, чувствительность и специфичность соответственно.

Например, рассмотрим население в 1 миллион человек. Если мы предположим, что 15% инфицированы SARS-CoV-2, то будет 150 000 инфицированных и 850 000 неинфицированных. В этом случае сделаем серологический тест для всех в населении. Тест имеет чувствительность 95%. Это означает, что тест точно даст положительный результат для 95% инфицированных людей - и специфичность 95%, что означает, что тест точно даст отрицательный результат для 95% неинфицированных людей. В этом примере мы также предполагаем, что наличие антител означает, что человек невосприимчив к инфекции SARS-CoV-2.

Исходя из этих условий, 185 000 человек будут иметь положительный результат теста; 142 500 будут истинно положительными, а 42 500 (23%) из них будут ложными.

Эти 42 500 человек получат положительный результат теста, но по-прежнему будут подвержены инфекции SARS-CoV-2. Таким образом, из 185 000 человек, считающих себя защищенными, 23% остаются уязвимыми.

Кроме того, 815 000 человек получат отрицательные результаты тестов, но 7 500 из них будут ложноотрицательными; это люди, которые, как считается, не имеют антител, но которые, возможно, уже были инфицированы и избавились от болезни. В целом, хотя общая точность высока (95%), и 95% всего населения получат точные результаты, как положительные, так и отрицательные, функционально из тех, у кого есть положительный результат теста и предполагаемый иммунитет, 23% все еще подвержены этому заболеванию – почти каждый четвертый. Только 77% положительных результатов будут точными – это прогностическая ценность положительного результата теста. Понимание ограничений теста имеет решающее значение для использования его в качестве инструмента для принятия политических или операционных решений.

Таким образом, последствия ошибок тестирования - ложноположительный или ложноотрицательный тесты не эквивалентны. Ложноположительный результат может помешать человеку вернуться к работе; ложноотрицательный может привести к цепочке эпидемии.

Также стоит помимо, непосредственно, тестирования отметить одну из значительных частей влияющих на точность результата, а именно, временной промежуток взятия теста. В зависимости от времени мы можем получить как очень точный результат (100%), так и не очень точный (<50%) пользуясь одним и тем же тестом.

Результаты тестирования изначально находятся в лабораториях или других медицинских учреждениях, поэтому будет резонно перейти непосредственно к ним.

Существует огромное количество лабораторий, как частных, так и государственных. По умолчанию доверительная доля больше у государственных лабораторий, так как их контролирует государство соответствующими механизмами. В случае же с частными, всё может быть немного иначе, об этом стоит помнить, однако их доверительную долю также можно поднять, добавив механизм проверки результатов с некоторой периодичностью. Помимо доверительной доли по умолчанию, стоит рассмотреть точность тестов, проводимых лабораторией, в случае если 100% тестов были выполнены с маленькой точностью (<50-60%), не следует сильно доверять результатам, несмотря на то что лаборатория выполнила всё по технологии и нигде не ошиблась. После лабораторий информацию о тестах собирает непосредственно государство.

ЛИТЕРАТУРА REFERENCES

[1] Download COVID-19 datasets “European Centre for Disease Prevention and Control” [Online]. Available: https://www.ecdc.europa.eu/en/covid-19/data

[2] COVID-19 resources “The Institute for Health Metrics and Evaluation (IHME)” [Online]. Available: http://www.healthdata.org/covid

[3] Coronavirus (COVID-19) Testing “Our World in Data” [Online]. Available: https://ourworldindata.org/coronavirus-testing

[4] Hannah Ritchie, Esteban Ortiz-Ospina, Diana Beltekian, Edouard Mathieu, Joe Hasell, Bobbie Macdonald, Charlie Giattino, and Max Roser. A cross-country database of COVID-19 testing. Sci Data 7, 345 (2020)

[5] COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University" url: https://github.com/CSSEGISandData/COVID-19

[6] Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Inf Dis. 20(5):533-534. doi: 10.1016/S1473-3099(20)30120-1

[7] Коронавірусна інфекція COVID-19. “Центр громадського здоров’я Міністерства охорони здоров’я України” url: https://phc.org.ua/kontrol-zakhvoryuvan/inshi-infekciyni-zakhvoryuvannya/koronavirusna-infekciya-covid-19

[8] Аналітичні панелі та відкриті дані. Відкриті дані. “Кабінет Міністрів України” url: https://covid19.gov.ua/analitichni-paneli-dashbordy