Balancing Precision and Recall for Cost-effective Epidemic Containment

Stefano Giovanni Rizzo
strizzo@hbku.edu.qa
Qatar Computing Research Institute
Doha, Qatar

ABSTRACT

In this paper we focus on the problem of containing an epidemic outbreak by selectively limiting the mobility of individuals. In a viral epidemic, such as the COVID-19 pandemic that we are currently facing, governments can enforce restrictions on a global or individual level, for example they may impose a country lockdown or they may quarantine only the suspected cases. We propose a solution that combines a high-recall model with a high-precision model, to select more or less individuals for isolation depending on the current infection rate, with the goal of minimizing both the number of interventions and the number of resulting infected cases. The solution placed third in the PAPW2020 Workshop Challenge on Mobility Intervention for Epidemics.

KEYWORDS

gradient boosting machine, prescriptive analytics, recall, epidemic

1 INTRODUCTION

In an effort of contrasting an epidemic outbreak, different strategy for human mobility restrictions can be applied. An optimal strategy for intervention policies is a strategy that, at the same time, minimizes both the spread of infections and the interventions on mobility.

In this context, the predicted fraction of infected population, or, on an individual scale, the predicted risk of infection based on contact tracing, should drive the direction and the extent of mobility restrictions, a problem that fits well in the paradigm of Prescriptive Analytics. In Prescriptive Analytics, predictive models and control models are combined to optimize a joint objective.

In the specific problem of managing mobility restrictions through prediction and control, the two costs of infections and interventions have strong ties with the concepts of precision and recall. Suppose that we aim at predicting and thus isolating the individuals who are infected but not yet discovered. Being bound by the precision/recall tradeoff, we could have a high precision but low recall prediction, intervening on few right individuals while many cases will keep

spreading, or we could have a high recall, low precision predictions, in which case the epidemic will quickly stop but with the cost of many unneeded interventions. This is better illustrated in Figure 1. Moreover, if the infection rate is very high, precise contact tracing is not a viable solution [1], leaving a full lockdown as the only option, that would be the equivalent of using a very high recall model.
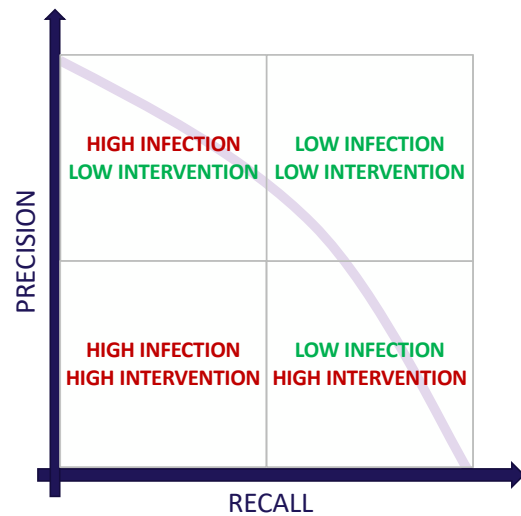

Figure 1: Example of the relation between precision, recall, infection cost and intervention cost in prescriptive analytics for mobility interventions in epidemics.

In this paper we propose an approach that takes advantage of two predictive models, one with high precision, one with high recall, to balance the intervention actions with the aim of balancing the two costs. We show that the current infection rate (or in epidemiological terms, the effective reproduction number R_t), should drive the minimum recall needed to stop the epidemic. Through experiments on the epidemic simulator of the *PAPW 2020 Challenge on Mobility Intervention for Epidemics*, we show how a single predictive models leads to a high cost in interventions or in infections, while the proposed approach obtain a substantial improvement in reducing both. The proposed approach has reached the third place in the challenge.

2 BACKGROUND

We hereby describe how the infection and intervention cost can be measured, and how the dynamics of infections and interventions in a real epidemic scenario have been modeled in the simulator of

the *PAPW 2020 Challenge on Mobility Intervention for Epidemics*[1].

Cost. Following a simplified version of the definitions and design used in the aforementioned challenge model, we formally define this dual cost as following. Given the cumulative number of infected cases in a 60 days episode I, and an isolation count $Q = \lambda_i N_{isolated}$, where $N_{isolated}$ is the count of person-days that sums each day that each person has been isolated, the overall cost of the strategy is:

$$cost = e^{\frac{I}{\theta_I}} + e^{\frac{Q}{\theta_Q}} \quad (1)$$

In this definition, $\lambda_i, \theta_I, \theta_Q$ are parameters of the cost, set in the challenge as $\lambda_i = 0.5$, $\theta_I = 500$ and $\theta_Q = 10000$.

Infection stages. An individual's health state can go through the following stages, in order:

- Stage 1: Susceptible. The individual has no infection therefore is liable to be infected. When infected from a contagious person, the individual moves to Stage 2.
- Stage 2: Pre-symptomatic. The individual is infected and contagious, but not yet detected. The infection is not observable. After an incubation period that can randomly go from 1 to 5 days, the individual moves to Stage 3.
- Stage 3: Symptomatic. The individual is infected and contagious, the infection is detected and observable.
- Stage 4: Symptomatic critical. The individual is infected, contagious and with critical health condition.
- Stage 5: Recovered. The individual is recovered, non-contagious and cannot be infected again. An individual self-recover after a period that can randomly go from 15 do 30 days.

Interventions. At the start of each day, the policy can decide whether to isolate an individual or not until the next day. When individuals are isolated, they cannot be infected, if they are in Stage 1, or infect anyone, if they are in one of the infected stages. We omit details on additional types of interventions, provided in the original simulator of the *PAPW '20 Challenge*, as these interventions are not used in the proposed approach.

Scenarios. We consider 5 different epidemic scenarios. In all the scenarios, 10000 people are simulated for 60 days. On weekdays, individuals spend 1 to 3 hours in their residential areas, then they move to work areas for 7 to 10 hours. After work, they may visit a commercial area with a certain probability and stay there 1 to 2 hours. At any time and place, a free susceptible individuals can get infected from a contagious person from the set of their work or residential acquaintances, with probability $P_c = 0.025$, or they could get infected from random strangers in the same area, with a lower probability $P_s = 0.005$. The above settings hold for the default Scenario 1, while the rest of the scenarios slightly deviates from it as following: Scenario 2 has double infection probability ($P_c = 0.05, P_s = 0.01$); Scenario 3 has 98 areas instead of the default 11; Scenario 4 starts on the first day with 300 infected individuals randomly picked; Scenario 5 has a larger range of start working

---

[1] https://hzw77-demo.readthedocs.io/en/round2/simulator_modeling.html

time, with individuals staying in the residential areas for 1 to 8 hours before moving to work.

Note that, during the first 20 days of each scenarios apart from Scenario 4, some people are randomly infected to simulate infections from outside the environment.

3 PROPOSED METHOD
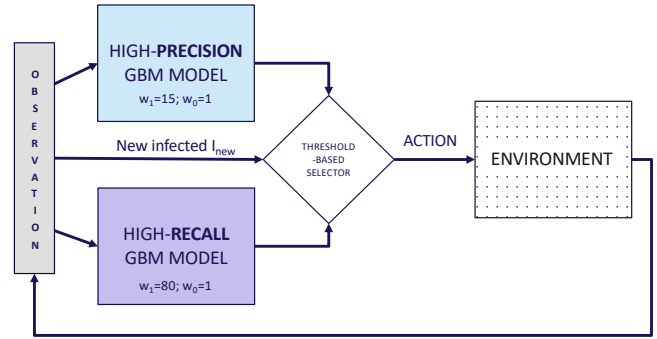
Figure 2: Full approach with high-recall and high-precision predictors.

In the proposed approach, two predictive models are trained, one with high precision, one with high recall, as shown in Figure 2. At each day, based on the current infection rate, one or the other model is used to predict whether each individual will be symptomatic in the next 5 days, and isolate all the predicted.

3.1 Prediction

The training set for the predictive models is built on a set of engineered features, each set representing the observation of an individual in environment. This set of features can be divided in local features and global features. The local features are specific for each individual:

- Ratio of acquaintances (residential and working), detected as symptomatic in the last 5 days.
- Probability of being infected based on number of times the individual was in the same time and place of a detected individual. (residential areas, working areas, commercial areas).
- Fraction of time spent under intervention in the last 5 days.
- State of infection of the individual.

The global features are the same for all the individual:

- Current time in episode as fraction of the whole episode.
- Day of the week, as a one hot encoding vector of size 7.
- Global infection state as new infected (binary), fraction of infected, fraction of new discovered in last step.
- Fraction of isolated individuals.
- Fraction of recovered individuals.

The predictive model is a Gradient Boosting Machine [2], optimizing a class-weighted log-loss defined as:

$$L_{w_0, w_1}(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} w_1 y_i \log(\hat{y}_i) + w_0(1 - y_i) \log(1 - \hat{y}_i) \quad (2)$$

with y the ground truth vector of size N, where $y_i = 1$ if the individual i was symptomatic in the following 5 days, $y_i = 0$ if the individual stayed in the susceptible stage during the subsequent 5 days, $\hat{y}$ is the vector of predictions, w_1 is the weight of the positive class error (infected), w_0 is the weight of the negative class error (not infected). A class-weighted loss is a common approach in cost-sensitive classification, usually applied in the presence of unbalanced samples [3]. Using the above loss, we can obtain a model with greater recall by increasing the value of positive weight class w_1 with respect to w_0.

3.2 Control

If an individual is detected as symptomatic, this is isolated by default, until found to be recovered. If an individual is susceptible, the related observation is used to get a prediction from the two models. Then, the fraction of new infected is used in a rule-based decision using a threshold γ and the current fraction of new infected I_{new}:

- If $I_{new} \leq \gamma$, the prediction from the high-precision model is used to for the isolation action.
- If $I_{new} > \gamma$, the prediction from the high-recall model is used until the end of the episode to contain the surge of infections. Additional, in the first 5 days after this event, all the individuals with a visit risk higher than a threshold are also isolated to further increase the recall.

By employing a higher recall prediction when the rate of infection is high, we ensure that the outbreak is contained. Generally speaking and under some assumptions (i.e. the intervention period is equal to the generation time), if $1 - \frac{1}{R_0} < recall$ the epidemic is contained. In Figure 3 we illustrate this by simulating the first two weeks of an epidemic with different reproduction numbers R_0 and 100 cases on the first day. With $R_0 = 2$, a system with a detection recall of 0.6 is sufficient to slow down and stop the epidemic after two weeks. If the same predictive model is applied when $R_0 = 2.6$, however, this is not enough to contain the outbreak. Applying a model with a recall of 0.8 to the same R_0, the epidemic is quickly contained within a week.

Tuning the recall of interventions depending on the infection rate allows us to effectively balance between recall and precision and, consequently, between the cost of extensive intervention and the cost of an epidemic outbreak.

4 RESULTS

We run three different methods for 20 episodes each: a model that only uses the high precision predictor, a model that only uses the high recall prediction, and finally the full approach presented in the paper. We choose to show the result for Scenario 2, as it is the most difficult scenario considering the higher probability of infections. The average cost and its breakdown in intervention and infection cost are shown in Table 1.

In this scenario, the high precision method obtains the worst result. Almost 3/4 of the cost is due to the high number of infections. Since by default we isolate infected individuals, the intervention cost is nevertheless conspicuous. Because the Scenario 2 has a higher probability of infection, a model with high recall is much more effective. For this model, the intervention cost prevail in the overall cost with a percentage of 99.83%. Finally, the proposed
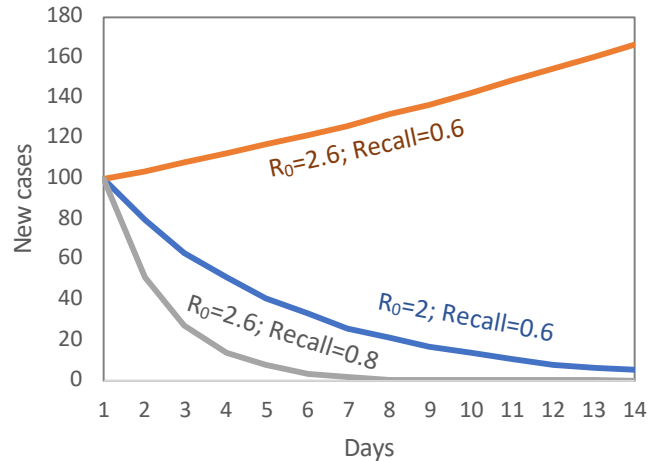
Figure 3: Simulation of epidemics with $R_0 = 2$ and $R_0 = 2.6$. A model with recall 0.6 is sufficient for $R_0 = 2$ but it results in an outbreak if $R_0 = 2.6$.

approach, which adaptively enable the high recall model when a super spread is taking place, obtains the best average result, with a cost that it is also mainly due to the interventions.

Method	Total cost	% infection cost	% intervention cost
High precision	49032.76	74.8%	25.2%
High recall	1179.20	0.17%	99.83%
Full approach	411.49	1.94%	98.06%

Table 1: Total cost for each method and percentages of cost for intervention and infection. Note that since we isolate infected people by default, higher infections will also result in additional interventions.

5 CONCLUSION AND FUTURE WORK

In this paper we presented a method that balances between high recall and high precision to detect and isolate infected individuals. We have shown that, depending on the current infection rate, a different level of recall is required to contain the epidemic. Taking advantage of this observation in the proposed solution, we obtain a lower overall cost and a better balance between infection costs and intervention costs. As a future work, we plan to design a prediction model that adaptively changes its sensitivity at inference time, depending on external parameters such as the global infection rate.

REFERENCES

[1] Luca Ferretti, Chris Wymant, Michelle Kendall, Lele Zhao, Anel Nurtay, Lucie Abeler-Dörner, Michael Parker, David Bonsall, and Christophe Fraser. 2020. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science* 368, 6491 (2020).
[2] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost: unbiased boosting with categorical features. In *Advances in neural information processing systems*. 6638–6648.
[3] Yanmin Sun, Mohamed S Kamel, Andrew KC Wong, and Yang Wang. 2007. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition* 40, 12 (2007), 3358–3378.