Các mô hình ước tính phù hợp khi dữ liệu còn hạn chế

Dự đoán tổng số ca mắc SARS-COV-2 tại Việt Nam theo dòng thời gian

Anh H. Ngo[1], Nam T. Hoang[2], and Khoi T. Nguyen[3]

[1]École Polytechnique, Institut Polytechnique de Paris, FRANCE
[2]Department of Mathematics and Computer Science, Beloit College, USA 53511
[3]Melbourne School of Engineering, The University of Melbourne, Parkville, Victoria, AUSTRALIA 3052

1 Lời mở đầu

1.1 Hiện trạng

Ở thời điểm hiện tại trong quá trình chống dịch tại Việt Nam, có rất nhiều dữ liệu đã thu thập được. Tuy nhiên, những dữ liệu này lại chưa được trích xuất một cách hoàn chỉnh, phần lớn là untidied data, có nghĩa là các dữ liệu lộn xộn, không thống nhất về hình thức cũng như nội dung. Từ vấn đề này, các nghiên cứu sinh ra nhu cầu sử dụng ít dữ liệu, nhưng có khả năng dự đoán nhanh các số liệu.

1.2 Yêu cầu

Để đáp ứng được nhu cầu nêu trên, cần phải thỏa mãn 3 yêu cầu chính:

- Tốc độ nhanh chóng
- Có độ chính xác cao
- Mang ý nghĩa thực tiễn

2 Giải pháp

Để đáp ứng được các yêu cầu đã nêu, hiện tại có các giải pháp như sau:

- Xây dựng mô hình mang số lượng biến số tối thiểu.
- Độ phức tạp mô hình đủ cao, nhưng không bị overfit (quá bám sát các dữ liệu đã có và tạo ra độ thiếu chính xác trong việc dự đoán).

⇒ Có thể áp dụng mô hình 1-dimensinoal (1 biến số), dựa trên số ca đã có từ dữ liệu của các ngày trước đó.

2.1 Ưu điểm

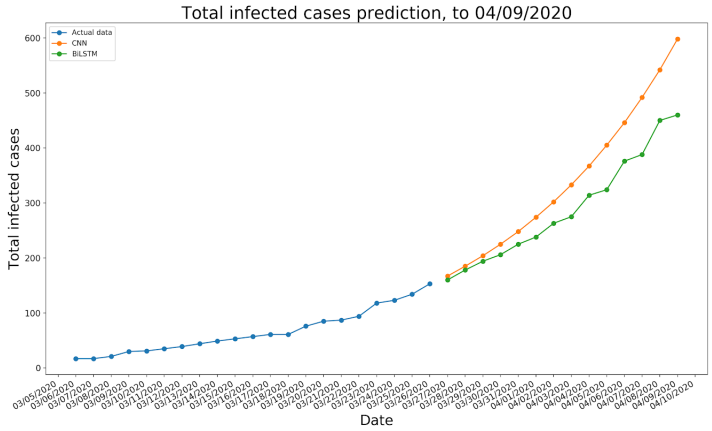
- Nhanh chóng
- Độ chính xác cao

⇒ **Đủ ý nghĩa để đưa ra giải pháp vĩ mô kịp thời.**

2.2 Nhược điểm

- Không mang ý nghĩa dịch tễ, do chỉ xác định được là nhiễm, chưa xác định được nguồn lây, F1,...
- Độ chính xác chưa đạt tuyệt đối. Sai số dao động ở mức 3-5%.

⇒ **Sẽ cải thiện được bằng nhiều dữ liệu biến khác nhau.**



Hình 1: Mô hình CNN và LSTM được áp dụng

3 Tiếp cận ban đầu

Các giải pháp để tạo nên bước đi đầu tiên sẽ được áp dụng từ các phương pháp trong 2 lĩnh vực:

3.1 Traditional Machine Learning

Lĩnh vực máy học truyền thống có thể được áp dụng thông qua những phương pháp sau:

- Convolutional Neural Network (CNN)
- Long Short-term Memory (LSTM)
- ARIMA family

3.2 Mô hình toán học

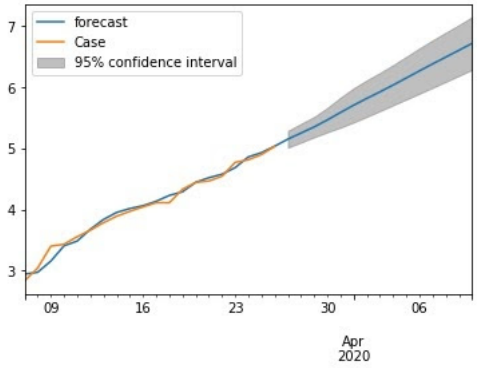
Để áp dụng bằng các mô hình, có thể sử dụng mô hình Grey và các bản mở rộng (Grey Model and extensions).

4 Mô hình Machine Learning truyền thống

4.1 Phân loại

Các mô hình Machine Learning truyền thống được áp dụng bao gồm:

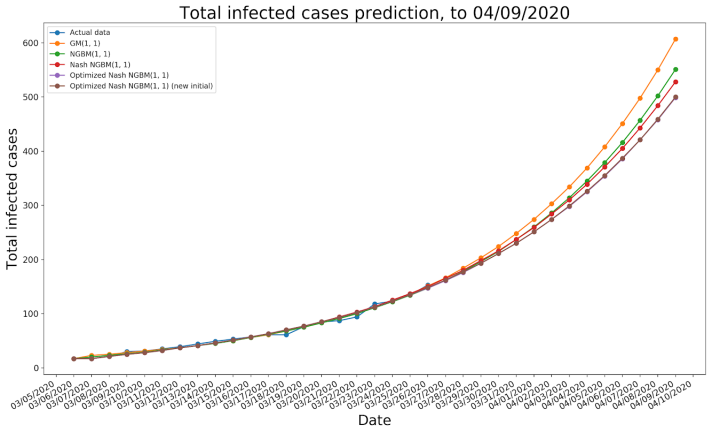
- Neural network
  - Dựa trên mạng neural thần kinh của con người.
  - CNN có nhiều mạng và neural ẩn để học tốt hơn các mạng máy học đơn thuần.
  - LSTM có nhiều cell (tế bào), hoạt động như bộ não của con người ,có chức năng “quên" để lọc data.
  - Biến thể Bidirectional LSTM (BiLSTM) có khả năng học từ hai chiều (input và output).
- ARIMA Family Là mô hình xác suất cơ bản, được dùng để benchmark các mô hình khác.



Hình 2: Mô hình ARIMA được áp dụng

Date	CNN	BiLSTM	ARIMA
03/27/2020	167	160	186
03/28/2020	185	178	204
03/29/2020	204	194	225
03/30/2020	225	206	250
03/31/2020	248	225	281
04/01/2020	274	238	314
04/02/2020	302	263	349
04/03/2020	333	275	386
04/04/2020	367	314	429
04/05/2020	405	324	479
04/06/2020	446	376	535
04/07/2020	492	388	595
04/08/2020	542	450	662
04/09/2020	598	460	735

Hình 3: Kết quả dự đoán của các mô hình Machine Learning truyền thống



Hình 4: Áp dụng Grey Systems

4.2 Kết quả

1. Không xuất hiện overfit khi sai số có xu hướng nhỏ dần.
2. Có sai số lớn vào những ngày nhất định (vào các ngày 10/03/2020 và 19/03/2020).
3. CNN dự đoán xu hướng tăng theo cấp số nhân, LSTM dự đoán có xu hướng theo ngày.
4. ARIMA dự đoán với sai số 5%, và lệch ± 2 ngày
5. Với data vào ngày 27/03/2020, dự đoán kết quả là 160-166 ca, kết quả chính xác là 163 ca, theo Bộ Y Tế (1).
6. Với data vào ngày 28/03/2020, dự đoán kết quả là 176 - 180 ca, kết quả chính xác là 174 ca, theo Bộ Y Tế.

5 Mô hình toán học

5.1 Mô tả Grey Systems & Extensions

- Ý tưởng dựa trên sự discretize theo thời gian của 1 phương trình vi phân.
- Đã được đề cập trong bài nghiên cứu của tác giả (2)
- Bao gồm: GM(1,1), NGBM(1,1), Optimized NGBM(1,1), Rolling optimized NGBM(1,1).

5.2 Kết quả

1. Không xuất hiện overfit khi sai số có xu hướng nhỏ dần.
2. Có sai số lớn vào những ngày nhất định (cụ thể ngày 4/3 và ngày 19/3).
3. Sai số trung bình nhỏ (nhỏ hơn 5% với những mô hình tối ưu hóa, ngày cuối cùng có sai số khoảng 3%).
4. Với data mới nhất hiện có (ngày 27/3/2020), dự đoán kết quả là 160-166 ca, kết quả chính xác ngày 27/3/2020 là 163 ca, theo Bộ Y Tế (1).

Date	GM(1, 1)	NGBM(1, 1)	Nash NGBM(1, 1)	Optimized Nash NGBM(1, 1)	Optimized Nash NGBM(1, 1) (new initial)
03/27/2020	166	162	165	161	162
03/28/2020	184	178	180	176	177
03/29/2020	203	196	198	193	193
03/30/2020	224	215	216	211	211
03/31/2020	248	237	237	230	230
04/01/2020	274	260	259	251	251
04/02/2020	303	286	284	274	274
04/03/2020	334	314	310	298	299
04/04/2020	369	345	339	325	326
04/05/2020	408	379	371	354	355
04/06/2020	451	416	405	386	387
04/07/2020	498	457	443	421	421
04/08/2020	550	502	484	458	459
04/09/2020	607	551	528	499	500

Hình 5: Kết quả dự đoán của các mô hình Grey Systems

6 Kết quả & Tổng kết

6.1 Dự đoán chung

Các dự đoán có thể phân loại thành 3 trường hợp theo thứ tự tình trạng xấu dần: Best Scenario (Tốt nhất), Arverage Scenario (Trung bình), Worst Scenario (Tệ nhất):

Ngày	**Best Scenario**	**Average Scenario**	**Worst Scenario**
31/03	225	230-235	240
09/04 (sau 2 tuần)	450-500	500-550	600

6.2 Hướng đi tiếp theo

- Mở rộng số lượng và phạm vi mô hình áp dụng.
- Sử dụng mô hình đa biến để các mô hình mang thêm ý nghĩa dịch tễ.
- Ngoài dự đoán số ca mắc của địa phương/cả nước, có thể dùng Machine Learning để dự đoán ca bệnh (xác suất phải dùng ICU, tử vong/hồi phục, ngày xuất viện,...)

Tài liệu

[1] https://ncov.moh.gov.vn/. *Thống kê tình hình dịch bệnh COVID-19*. Bộ Y Tế, Việt Nam, 2020.

[2] HA, Ngo, TN, Hoang. *A rolling optimized nonlinear grey bernoulli model ROMGBM(1,1) and application in predicting total 2019-nCoV infected cases*. 2020.