Vol.2020-MBL-96 No.25
Vol.2020-UBI-67 No.25
Vol.2020-CDS-28 No.25
Vol.2020-ASD-19 No.25
2020/9/30

勾配ブースティング決定木を用いた国内公開データにおける COVID-19 の重症化に影響する重要因子の解析

渡辺大智[1]　鏑木崇史[1]

概要：COVID-19 の感染拡大が世界的にも広がっており、2020 年 8 月現在、世界では 1900 万人の感染者がおり、73 万人以上の死亡が報告されている。COVID-19 については解明されていないこともまだ多く、その死亡率の高さと無症状患者の多さが予防を困難にしている。しかしながら、最近の研究によると死亡率や重症化患者の割合は国によって異なることが報告されている。例えば、入院患者の内、中国では 28%、イギリスでは 26%、アメリカでは 21%-24%の死亡が報告される一方で、日本では 7.8%にとどまっている。
本論文では、死亡率の違いを模索する目的で勾配ブースティング決定木を用いて公開されている患者データを解析する。まずは国内のデータを解析し、重症化につながる重要因子を推定した。その結果、70 歳以上から突然死亡率が高くなる傾向や、基礎疾患持ちの患者、さらに男性の方が死亡患者の割合が多いという傾向が得られた。また、無症状患者は症状がある患者と違い、広い地域に出現していることから、まだ患者が多く出ていない地域に対しての感染予防策や、死亡リスクの高い 70 歳以上で基礎疾患のある患者に対する警戒が必要であることが公開データからも示唆された。

キーワード：COVID−19，勾配ブースティング決定木

1. はじめに

中国の武漢から世界中に広がった COVID-19 は、2020 年 9 月現在、世界で 2600 万人以上の感染者がおり、86 万人以上の死者が報告されている。COVID-19 に感染したとされる患者 44000 人を分析した中国の報告では、高齢、心血管疾患、糖尿病、慢性呼吸器疾患、高血圧、がんが死亡リスクの増加と関連していることが示された[1]。また、46248 人の患者を含む 8 つの研究では、最も重篤な疾患を有する患者は高血圧（オッズ比 2.36（95%信頼区間 1.46～3.83））、呼吸器疾患（2.46（1.76～3.44））、心血管疾患（3.42（1.88～6.22））を有している可能性が高いことが示された。[2]。他の研究では、肥満と喫煙がリスクの増加と関連していた [3][4]。

以上のように、COVID-19 による死亡の危険因子は多くの国で報告されているが、日本ではそれらに焦点を当てた研究は少なく、症状の重症度に影響を与える因子を分析した研究はほとんど存在していない。そこで本研究では、国内の COVID-19 感染患者を無症状患者、要治療患者、死亡者の 3 つの段階に分け、それぞれを分類するときにおいてどのような因子がどの程度影響しているかを明らかにすることを目的とし、また、これらの結果からどのような感染予防策が効果的であるのかを考察した。これらの目的を達成するために，LightGBM を用いて、オープンデータセットから日本国内の 5842 人の患者データを解析した。

2. 実験

使用するデータセットの準備から分析までの流れは、次の通りである.

(1) 使用するデータセットと特徴量

本研究で使用する日本国内の患者データは、29 の特徴量と 23315 人分のデータがから構成される SIGNATE COVID-19 Case Dataset を元に、81 の特徴量と 5842 人分のデータに整形したものを使用した。具体的に、特徴量は性別、年齢、場所、基礎疾患、渡航歴、ダイヤモンド・プリンス号の搭乗歴、症状、職業、発症から公表までの日数、発症から確認までの日数、確認から公表までの日数、都道府県別症例番号、重症度を使用した（表 1）。なお表 1 には各特徴量の欠損値が含まれていない。

(2) データの前処理

重症度については、元のデータセットのうち、「軽度」と「中等度」を「要治療」に置き換え、患者全体の重症度を「無症状」、「要治療」、「死亡」の 3 つに分類したのち、それぞれ順に「0」「1」「2」と数値のクラスに変更した。職業については、252 種類の中からデータセットで最も頻度の高い 10 種類に職業を抽出した。また、データセットから「公表日から発症までの日数」「確認日から発症までの日数」「公表日から確認日までの日数」「公表日から確認日までの日数」という新たな特徴量を作成した。

データのサンプル数については、「無症状」が 919 件、「要治療」が 4765 件、「死亡」が 158 例と偏りがあるため、ダウンサンプリングを用いてモデルを作成した。すべてのクラスを 158 件にダウンサンプリングしモデルの学習を進めることでモデルの精度が大幅に向上した。

(3) 分析ツール・評価方法

データ分析は、Google Colaboratory 上にて LightGBM モデ

1 国際基督教大学
International Christian University

Vol.2020-MBL-96 No.25
Vol.2020-UBI-67 No.25
Vol.2020-CDS-28 No.25
Vol.2020-ASD-19 No.25
2020/9/30

ル(n_estimators=100, n_stopping_rounds=100)を用いて解析した。データセットのうち 20%をテストデータとし、モデルの予測精度は、正解率、適合率、再現率、f 値と混同行列を用いて評価した。特徴量の重要度はジニ不純度を用いて算出した。

表 1　データセットの内訳(N=5842：無症状者=919, 要治療者 =4765, 死亡者=158)

特徴量		無症状	要治療	死亡
性別	女性	540	2173	54
	男性	360	2572	102
年齢	0 - 9	73	63	0
	10 - 19	69	97	0
	20-29	143	720	0
	30-39	100	668	1
	40 - 49	116	799	3
	50 - 59	112	851	5
	60 - 69	90	579	15
	70 - 79	84	481	48
	80 - 89	79	320	60
	90 - 99	39	145	23
基礎疾患	なし	104	1033	24
	あり	20	236	43
海外渡航歴	なし	375	2353	69
	あり	38	192	3
ダイヤモンド・プリンス号の搭乗歴	なし	389	2004	72
	あり	38	85	4
症状	発熱	7	1825	34
	肺炎	2	262	14
	咳	2	938	18
	咽頭痛	0	272	1
	意識障害	0	93	1
	なし	74	3	0
職業	会社員	93	1207	14
	無職	146	937	69
	自営業	15	221	3
	医療従事者	53	259	158
	接客業	3	81	2
	学生	83	143	0
	公務員	4	88	0
	アルバイト	10	94	1
	介護職	22	70	0
	施設入所者	0	0	0
発症から公表までの日数	0 – 51 日	33	4346	142
発症から確認までの日数	0 - 51 日	30	3185	104
確認から公開までの日数	0 - 48 日	811	3508	109
場所	50 種類	919	4765	158

Vol.2020-MBL-96 No.25
Vol.2020-UBI-67 No.25
Vol.2020-CDS-28 No.25
Vol.2020-ASD-19 No.25
2020/9/30

3. 結果

3.1 重要因子

モデルの正解率は 0.86 であった。また、適合率、再現率、f 値のいずれも一貫して高いことから，モデルはデータを正しく分類できると考えられる（図 1）.

日本のデータセットでは，都道府県別症例番号，年齢，公表から発症までの日数が患者の分類に重要であるが，職業，症状，都道府県はほとんど重要ではなかった（表２）。

3.2 因子ごとの詳細

重要因子の詳細に注目すると、都道府県別症例番号が大きいほど、また、年齢が高いほどより重症度が深刻化しやすいと考えられた（図２図３）。特に年齢と重症度の関係においては、70 歳を超えると死亡する確率が格段に上がるということが明らかになった（図 3）。

場所に注目すると、要治療者や死亡者は同じような地域にまとまって出現するのに対し、無症状患者は広い範囲で出現していることが明らかになった（図 4）。

その他の特徴量をみると、死亡患者は他の重症度の場合と違い、女性よりも男性の方が多いということや、死亡患者は他の重症度の場合と違い、基礎疾患のある人の方がない人よりも突出して多いということが明らかになった（図 5）。重要度の大きい「発症から公表までの日数」、「発症から確認までの日数」、「海外渡航歴」などに関しては、重症度とどのような関係があるのか内訳からは明らかにできなかった（図 5）。

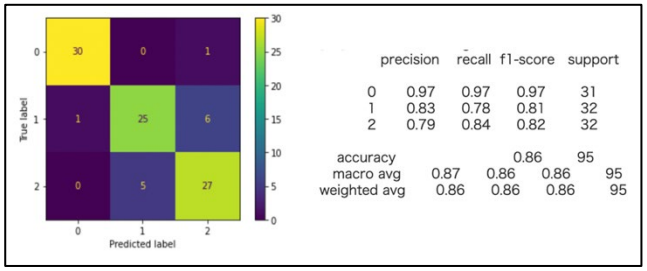

図 1　モデルの精度

表 2　重要度の高い特徴量

特徴量の重要度 (学習データ数=393、ジニ不純度≒0.67)	
都道府県別症例番号	277 (24.6%)
年齢	216 (19.2%)
発症から公表までの日数	154 (13.7%)
場所	89 (7.9%)
発症から確認までの日数	79 (7.0%)
海外渡航歴	76 (6.8%)
女性	70 (6.2%)
基礎疾患	45 (4.0%)
男性	29 (2.5%)
無職	18 (1.6%)
確定から公表までの日数	12 (1.1%)
発熱	11 (1.0%)
ダイヤモンド・プリンス号の搭乗歴	11 (1.0%)
神奈川県での受診	11 (1.0%)
咳	8 (0.7%)
大阪府での受診	4 (0.4%)
会社員	3 (0.3%)

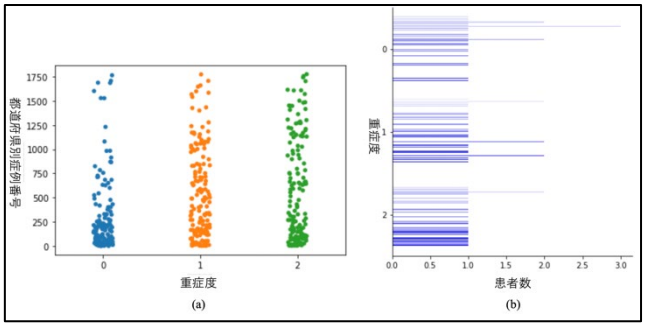

図 2　都道府県別症例番号と重症度の比較

無症状者の平均値=288.7、中央値=150（N=157）。要治療者の平均値=542.8、中央値=433（N=158）。死亡者の平均値=556.8、中央値=337（N=158）

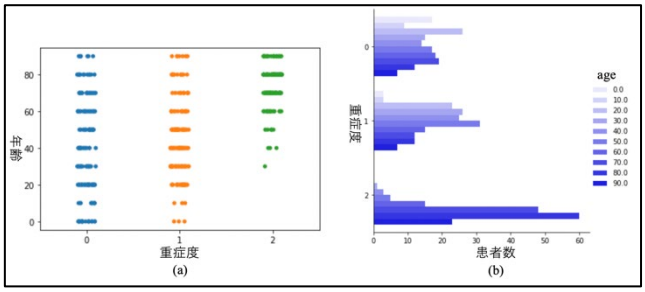

図 3　年齢と重症度の比較

無症状者の平均値=42.0、中央値=40（N=154）。要治療者の平均値=45.5、中央値=40（N=157）。死亡者の平均値=74.4、中央値=80（N=155）

Vol.2020-MBL-96 No.25
Vol.2020-UBI-67 No.25
Vol.2020-CDS-28 No.25
Vol.2020-ASD-19 No.25
2020/9/30

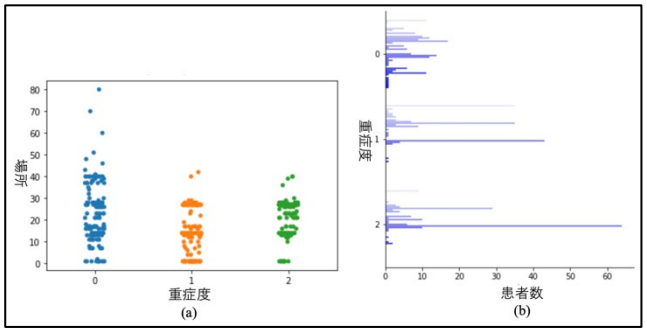

図 4　場所と重症度の比較

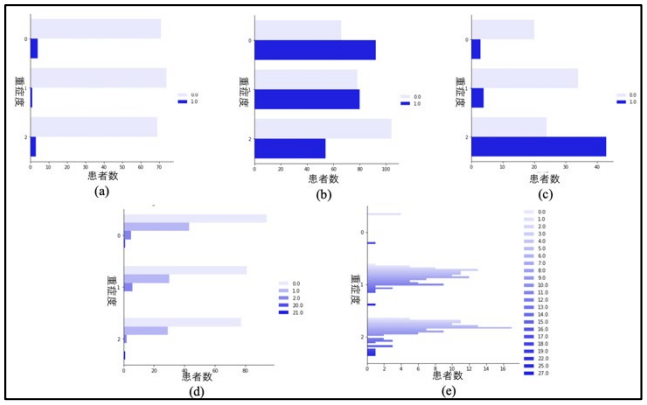

図 5　その他の特徴量と重症度の比較

(a)海外渡航歴と重症度の比較、(b)女性と重症度の比較、(c)基礎疾患と重症度の比較、(d)確定から公表までの日数と重症度の比較、(e)発症から確定までの日数と重症度の比較

4. 考察・展望

この研究は、公開されたオープンデータセットのうち必要な情報が含まれている患者データのみを対象に分析したため、データが公開されていない患者や十分な情報を持っていない患者を扱わなかかった。そのため、現在公開されている国内の COVID-19 感染患者数に比べて非常に少ないサンプル数しか分析対象にできていない。また、対象データセットには多くの欠損値が含まれていた。これらのことから、よりモデルの精度を上げるには、より多くの整形された感染患者のデータが必要になると考えられる。

死亡した症例の場合のみ、男性が女性を上回っていたことや、基礎疾患がある人の方が死亡リスクを格段にあげる因果関係は今回の研究では明らかに出来なかった。

5. 結論

都道府県別症例番号が大きい人ほど、要治療や死亡など重症化しやすくなるという傾向から、特定の地域における感染の拡大率と重症化率には相関関係があると考えれた。このことから、特定の地域における感染拡大を防ぐことは、要治療者や死亡者を減らすために役立つと考えられた。

今回の研究では、70 歳未満では症状があっても死亡する確率が低いのに対し、70 歳以上では感染して死亡する確率が特に高くなるということが明らかになった。また、COVID-19 感染患者では、男性であることや、基礎疾患を持っていることも死亡確率を高めている。これらの結果から，70 歳以上の人や基礎疾患を持つ人、そのうち特に男性のに対しては特別な感染予防対策が必要であることが示唆された。

また、重症化した患者は特定の範囲に出現するのに対し、無症状の患者は広範囲に出現していることから、まだ症状のある感染者が確認されていない地域でも無症状感染者がいる可能性は高いことから感染予防対策が十分に必要であるということが明らかになった。

本研究では、COVID-19 感染者の重症度を変化させる重要な要因を特定し、リスク要因を考慮した感染予防策を提言することができた。しかし、COVID-19 患者の情報が十分に公開されていない点や、収集した患者データの特徴量が統一されてなく、分析する際に多くの欠損値を含めてなくてはいけないという点から、モデルの精度に対する不満点は残った。より統一化されたデータの収集とそれらの公開は、より効果的な感染予防策を考察する上に非常に重要であると思われる。

謝辞

本研究は JSPS 科研費 16K16392 の助成を受けたものです。

参考文献

[1] Wu Z, McGoogan JM. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72314 cases from the Chinese Center for Disease Control and Prevention. JAMA2020. doi:10.1001/jama.2020.2648 pmid:32091533

[2] Yang J, Zheng Y, Gou X, et al. Prevalence of comorbidities in the novel Wuhan coronavirus (COVID-19) infection: a systematic review and meta-analysis. Int J Infect Dis2020;S1201-9712(20)30136-3. doi:10.1016/j.ijid.2020.03.017 pmid:32173574

[3] Huang R, Zhu L, Xue L, et al. Clinical findings of patients with coronavirus disease 2019 in Jiangsu Province, China: a retrospective, multi-center study. 2020

[4] Wang D, Hu B, Hu C, et al. Clinical characteristics of 138 hospitalized patients with 2019 Novel Coronavirus–Infected pneumonia in Wuhan, China. JAMA2020;323:1061-9. doi:10.1001/jama.2020.1585 pmid:32031570