Using data mining for analysing clinical and demographic risk factors of Covid-19 severe cases in Brazil

Dr. Renato Vimieiro
Postal address: Av. Antônio Carlos, 6627 - Prédio do ICEx Pampulha, Belo Horizonte, Minas Gerais, Brasil – CEP: 31270-901
Email:rvimieiro@dcc.ufmg.br
Phone: +55 (31) 3409 5860
Affiliation: Departamento de Ciência da Computação (DCC) – Universidade Federal de Minas Gerais (UFMG)

Abstract. The Coronavirus disease 2019 (COVID-19) was first detected in China in December 2019, and, in a few months, the disease got pandemic proportion. Risk factors related to the progression and outcome of the disease are still unclear. Moreover, clinical aspects of patients can differ between societies, and other demographic elements may impact survival response. A better characterisation of local manifestation of COVID-19 is crucial to a better general understanding of the disease, and thus to improve treatment decisions and health systems' management. We propose the investigation of risk factors of COVID-19 confirmed Brazilian patients, based on clinical and demographic information from over a million records from Brazilian epidemiological open database, integrated with data from Brazilian Institute of Geography and Statistics. For this matter, we propose the use of Exceptional Model Mining – a Data Mining method for discovering subgroups with distinct behaviour – with the goal of discovering combined clinical/demographic factors of COVID-19 associated with disease progression and outcome.

Problem Statement and Research Goals

The Coronavirus disease 2019 (COVID-19), caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), was first detected in late December 2019, in the city of Wuhan, China. Since then, it has spread quickly nationwide, taking pandemic proportions. The high spread rate of SARS-CoV-2 has overloaded hospitals all around the world, and governments have struggled to manage the disease progression. Risk characterisation of the disease is essential to better assess the course of treatments and improve managements decisions for health systems. However, many medical and demographic aspects of the disease are still under investigation, and the factors that interfere in the disease prognosis are, still, unclear. This context motivates us to pose the following research question: *which are the factors related to the disease progression in Brazil?*

Therefore, we aim at identifying risk factors – medical and demographic aspects – that may be associated with severe cases of COVID-19 in Brazil. Hence, we propose a new technique that combines the rule-based Data Mining task of Exceptional Model Mining and statistical methods of Survival Analysis for discovering subgroups of patients presenting unusual survival behaviour. Thus, our goals are as follows:

- Discover combinations of clinical factors that may be related to distinct – better or worse – survival response of COVID-19.
- Assess the influence of demographic aspects combined with clinical factors on the disease outcome in Brazilian patients.
- Provide understandable and straightforward characterisation of subgroups of patients presenting distinct survival behaviour based on clinical and demographic aspects.

Work Description and Expected Outcomes

We are developing an Exceptional Model Mining (EMM) framework that employs Ant-Colony Optimisation metaheuristics in the search for subgroups of patients that present survival functions which are statistically different from the overall survival of COVID-19 patients. We propose to use the Kaplan-Meier survival curves as characterisation of the survival behaviour, and the Logrank statistical test to assess differences in survival responses. The resultant rule model of EMM framework comprises a set of rules that describe subgroups with exceptional survival behaviour. Each rule is composed by conjunctions of features-values associated with statistically significant survival functions. In that way, the algorithm can provide identification of data characteristics and feature's interactions that interfere in survival experience.

The proposed algorithm is a general framework for mining subgroups with exceptional survival behaviour. We have implemented and tested an initial version of the algorithm in different survival databases. The tests' results showed that the algorithm is able to discover significant subgroups and to identify data characteristics that interfere in survival experience. Besides, we have performed initial processing on the Brazilian epidemiological open database of the severe acute respiratory syndrome. In the processing, we extracted over a million records related to COVID-19 confirmed patients in Brazil (until July 14th, 2020) and generated 47 clinical features related to the most observed symptoms and comorbidities. An initial exploratory analysis showed results similar to findings reported in the literature (Huang, et al., 2020) (Richardson, et al., 2020) (Song, et al., 2020).

For the goals intended in this project proposal, we plan to expand this initial work in two directions: (1) further development of the EMM framework; and (2) further processing and expansion of the data set. In order to refine the algorithm results, we plan to add the functionality of coping with numerical features, improve its performance on large and high-dimensional data, and adjust the meta-parameters of the pattern search. In addition, it is necessary to tackle redundancy of the discovery patterns and false statistical discoveries so that we can achieve simple, concise, and reliable results. As for the data to be investigated, we plan to expand the initially processed dataset of symptoms and comorbidities with further medical and demographic information. The additional medical data includes administrated medicines, results from chest x-rays, use of mechanical ventilators, and viral respiratory co-infection. The demographic aspects include age, ethnicity, municipality, and other census information available by the Brazilian Institute of Geography and Statistics (IBGE) – such as social indicators and statistics; municipality profile and economy; family income; education and professional qualification; and health system profile and economy. Therefore, in addition to code development, we will collect the data from different public domains, perform data cleaning and preprocessing, and expand the exploratory data analysis. Lastly, we will employ the proposed algorithm to learn from the data patterns associated with COVID-19 progression.

By employing our proposed Exceptional Model Mining framework in the investigation of risk factors of COVID-19, we expect that the algorithm will be able to retrieve unknown interactions between clinical and demographic aspects that are associated with the disease prognosis. Additionally, we expect our framework to be a contribution to the machine learning community as a tool for discovering and characterising subgroups with exceptional survival behaviour, not only for COVID-19 investigation but also for other domains. Finally, we intend to make our results available to the public, and we expect to contribute with a better understanding on the overall risk factors of COVID-19, as well as contributing to a better characterisation of the disease progression in Brazil.

Prior Work

It is known that the infection caused by SARS-CoV-2 affects mainly the respiratory capacity of patients and present a higher mortality rate in older ages. Clinical studies (Huang, et al., 2020) (Richardson, et al., 2020) (Song, et al., 2020) conducted in China and USA have reported the main symptoms and comorbidities observed in COVID-19 patients, indicating that the chance of mortality increases in the presence of coexisting medical conditions. Some studies (Zhou, et al., 2020) (Zheng, et al., 2020) (Schnake-Mahl, et al., 2020) have attempted to describe risk factors for the progression of COVID-19, mainly with data of patients from China and USA. However, the risk characterisation for distinguishing patients with higher chances of complications is still unclear. Furthermore, (Dowd, et al., 2020) highlight the role of demography for understanding differences in cross-country fatality and the impact of the pandemic on different populations, and, in this sense, (Nepomuceno, et al., 2020) highlight that different distributions of demographic aspects across diverse populations may reflect differently on fatality risk. When compared to China and, mainly, to Europe and USA (great focal points of COVID-19), Brazil presents very distinct demographic characteristics. Social and economic aspects may accentuate risk factors that are not significant in other countries, or even in different regions of Brazil. However, to our knowledge, there is still no report on risk factors tailored to Brazilian patients.

Exceptional Model Mining (EMM) (Leman, et al., 2008) is a data mining task at the intersection of predictive and descriptive perspectives. It is concerned with describing patterns related to a property of interest, which consists of any sort of model fitted to attribute targets – in our proposal, the Kaplan-Meier (KM) survival function. Given the KM model, the EMM task searches for subgroups of the data for which the model fitted to the subgroup differs substantially from the respective model fitted to the whole data. This perspective comprehends the data as a potential composition of different data subsets that present distinct survival behaviour. The EMM task has been applied in the study of different behaviours in different areas, e.g. (Lemmerich, et al., 2016) (Du, et al., 2020)

(Belfodil, et al., 2020). Its concept of property of interest makes it possible to use the survival model as target and, therefore, capture important information of the patients' survival experience in the analysis of exceptional survival behaviours.

However, to the best of our knowledge, no work on literature explores the use of EMM – or any other supervised descriptive pattern mining task – to uncover subgroups with unusual survival behaviour. Most of the machine learn and data mining approaches to Survival Analysis (Wang, et al., 2017) aim either to predict survival distribution or to classify new observations, rather than understand the characteristics that delineate subgroups with distinct survival behaviour. Additionally, our proposed algorithm is the first approach for EMM task that explores a bio-inspired meta-heuristic to optimize the pattern search. Finally, our EMM rule-based model comprises individual local patterns related to exceptional survival functions. In addition, the resultant rule-based model provides clear information on features' interactions and features' associations with survival response, posing a great tool for helping in the understanding of COVID-19 characteristics and in the support of medical and health-managements decisions.

References

Belfodil, A., Cazalens, S., Lamarre, P. & Plantevit, M., 2020. Identifying exceptional (dis)agreement between groups. *Data Mining and Knowledge Discovery,* 34(2), pp. 394-442.

Dowd, J. B. et al., 2020. Demographic science aids in understanding the spread and fatality rates of COVID-19. *Proceedings of the National Academy of Sciences of the United States of America,* 117(18), pp. 9696-9698.

Du, X., Pei, Y., Duivesteijn, W. & Pechenizkiy, M., 2020. Exceptional spatio-temporal behavior mining through Bayesian non-parametric modeling. *Data Mining and Knowledge Discovery.*

Huang, C. et al., 2020. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet,* 395(10223), pp. 497-506.

Leman, D., Feelders, A. & Knobbe, A., 2008. *Exceptional model mining.* s.l., s.n., pp. 1-16.

Lemmerich, F. et al., 2016. Mining subgroups with exceptional transition behavior. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* Volume 13-17-Augu, pp. 965-974.

Nepomuceno, M. R. et al., 2020. Besides population age structure, health and other demographic factors can contribute to understanding the COVID-19 burden. *Proceedings of the National Academy of Sciences of the United States of America,* 117(25), pp. 13881-13883.

Richardson, S. et al., 2020. Presenting Characteristics, Comorbidities, and Outcomes among 5700 Patients Hospitalized with COVID-19 in the New York City Area. *JAMA - Journal of the American Medical Association,* 323(20), pp. 2052-2059.

Schnake-Mahl, A. S., Carty, M. G., Sierra, G. & Ajayi, T., 2020. Identifying Patients with Increased Risk of Severe Covid-19 Complications: Building an Actionable Rules-Based Model for Care Teams. *NEJM Catalyst Innovations in Care Delivery,* 1(3).

Song, F. et al., 2020. Emerging 2019 novel coronavirus (2019-nCoV) pneumonia. *Radiology,* 295(1), pp. 210-217.

Wang, P., Li, Y. & Reddy, C. K., 2017. Machine learning for survival analysis: A survey. *ACM Computing Surveys,* 51(6), pp. 1-39.

Zheng, Z. et al., 2020. Risk factors of critical & mortal COVID-19 cases: A systematic literature review and meta-analysis. *Journal of Infection,* Issue 568.

Zhou, F. et al., 2020. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet,* 395(10229), pp. 1054-1062.

Data Policy

This work is intended to contribute to the research and understanding of the COVID-19. Therefore, we intend to publish its findings and make available the processed data set, as well as all necessary code for processing.

CURRICULUM VITAE – DR. RENATO VIMEIRO

EDUCATION AND QUALIFICATIONS

- **Post-Doc (2014)** – The University of Newcastle, Australia – 2014
- **PhD (2012) – Computer Science** – The University of Newcastle, Australia – 2012
- **Master (2007) – Computer Science** – Universidade Federal de Minas Gerais – 2007
- **Graduation (2004) – Computer Science –** Pontifícia Universidade Católica de Minas Gerais

CAREER SUMMARY

- **2019 – Present: Lecturer (Professor Adjunto)**, Universidade Federal de Minas Gerais
- **2014 – 2019: Lecturer (Professor Adjunto)**, Universidade Federal de Pernambuco
- **2012 – 2014: Research Associate**, The University of Newcastle/Hunter Medical Research Institute, Australia

PUBLICATIONS AND RESEARCH (MOST RELEVANT)

Articles in Journals

- Lucas, T., Silva, T.C.P.B., **Vimieiro, R.**, & Ludermir, T.B. (2017). A new evolutionary algorithm for mining top-k discriminative patterns in high dimensional data. Appl. Soft Comput. 59, 487–499. DOI: j.asoc.2017.05.048 (Qualis A1)
- Milioli, H.H., **Vimieiro, R.**, Tishchenko, I., Riveros, C., Berretta, R., & Moscato, P. (2016). Iteratively refining breast cancer intrinsic subtypes in the METABRIC dataset. BioData Mining 9, 2. https://doi.org/10.1186/s13040-015-0078-9 (Qualis A1)
- Milioli, H. H., **Vimieiro, R.**, Riveros, C., Tishchenko, I., Berretta, R., & Moscato, P. (2015). The Discovery of Novel Biomarkers Improves Breast Cancer Intrinsic Subtype Prediction and Reconciles the Labels in the METABRIC Data Set. PloS one, 10(7), e0129711. https://doi.org/10.1371/journal.pone.0129711 (Qualis A1)
- **Vimieiro, R.**, & Moscato, P. (2014). Disclosed: An efficient depth-first, top-down algorithm for mining disjunctive closed itemsets in high-dimensional data. Inf. Sci., 280, 171-187. (Qualis A1)
- **Vimieiro, R.**, & Moscato, P. (2014). A new method for mining disjunctive emerging patterns in high-dimensional datasets using hypergraphs. Inf. Syst., 40, 1-10. (Qualis A1)
- Arefin, A. S., **Vimieiro, R.**, Riveros, C., Craig, H., & Moscato, P. (2014). An information theoretic clustering approach for unveiling authorship affinities in Shakespearean era plays and poems. PloS one, 9(10), e111445. https://doi.org/10.1371/journal.pone.0111445 (Qualis A1)
- **Vimieiro, R.**, & Moscato, P. (2012). Mining disjunctive minimal generators with TitanicOR. Expert Syst. Appl., 39, 8228-8238. (Qualis A1)

Conference articles

- Lucas, T., **Vimieiro, R.**, & Ludermir, T.B. (2018). SSDP+: A Diverse and More Informative Subgroup Discovery Approach for High Dimensional Data. 2018 IEEE Congress on Evolutionary Computation (CEC), 1-8. (Qualis A1)

Book chapters

- Riveros, C., **Vimieiro, R.**, Holliday, E. G., Oldmeadow, C., Wang, J. J., Mitchell, P., Attia, J., Scott, R. J., & Moscato, P. A. (2015). Identification of genome-wide SNP-SNP and SNP-clinical Boolean interactions in age-related macular degeneration. Methods in molecular biology (Clifton, N.J.), 1253, 217–255. https://doi.org/10.1007/978-1-4939-2155-3_12

Quantitative Indicators

- Publications in journals with selective editorial policy: **8**
- Book chapters: **1**
- Doctoral theses oriented and defended: **1**
- Master's dissertations oriented and defended: **3**

Orientations in Progress

- Luana da Costa Faria. Masters Degree in Computer Science - Universidade Federal de Minas Gerais.
- Cecília Regina Oliveira de Assis. Masters Degree in Computer Science - Universidade Federal de Minas Gerais.
- Juliana Barcellos Mattos. Masters Degree in Computer Science - Universidade Federal de Pernambuco. Supervisor.
- Luís Fred Gonçalves de Souza. Masters Degree in Computer Science - Universidade Federal de Pernambuco. Supervisor.
- Andréa Brandão Duque. Masters Degree in Computer Science - Universidade Federal de Pernambuco. Supervisor.

List of Research Funding Under Responsibility Of The Researcher

Present

- **Edital Universal 2018**: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Process: 422346/2018-7, Title: *Investigação do uso de computação evolucionária para mineração de padrões discriminativos em bases de dados de alta dimensionalidade*, Duration: 2019-02-18 to 2022-02-28, Funding: R$20,000.00.

Concluded

- Edital Facepe/CNRS: Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco. Process: APQ1529-1.03/15. Title: On the development of a new method for analyzing multi-source biomedical data using multi-view clustering and pattern mining. Funding: R$73,200.00.
- Edital Facepe/PPP: Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco. Process: APQ0585-1.03/14. Title: *Mineração de padrões discriminativos em grandes volumes de dados biomédicos*. Funding: R$39,200.00.

Links for Academic Profile

- ORCID: https://orcid.org/0000-0002-7911-2456
- MyResearcherID (ISI): https://publons.com/researcher/G-8109-2013/
- MyCitations (Google Scholar): https://scholar.google.com.br/citations?hl=en&user=k1fUR_wAAAAJ