Extracting Association Rules from Blood Count Values of COVID-19 clinical Data

Christina Morgenstern
christinamorgenstern@lewisu.edu
DATA-51000-002, 20
Data Mining and Analytics
Lewis University

I. INTRODUCTION

The Coronavirus disease 2019 (COVID-19) caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has caused a pandemic in 2020 with more than 7 million people infected and more than 400,000 deaths [1]. First identified in Wuhan, China [2], this new virus spread across the globe mediated through close contact via droplets produced by sneezing, coughing and probably talking [3]. The symptoms caused by the disease are diverse with mild cases including fatigue, cough, shortness of breath and fever to acute respiratory distress syndrome with multi-organ failure and septic shock [4]. The success of the virus is also attributed to leaving many infected people without symptoms [5]. So far, there is no vaccine or specific antiviral treatment for COVID-19 [3] which left many countries assign lockdown, social distancing and hygiene measures to protect their people and national health institutions. While age is one of the major risk factors for a severe prognosis of COVID-19, little is known about other parameters involved i.e. metabolic or genetic factors, that lead to severe outcomes of infected patients.

In this paper, data from blood count values of COVID-19 patients were used to find association rules between blood parameters and the prognosis of a SARS-CoV-2 infection. The aim of this study is to find associations between blood parameters that contribute to worsening of patient outcomes. Such rules could inform clinical personnel if a certain patient with COVID-19 will need admittance to intense care units (ICU). This information can help in planning for the availability of ICU beds.

The following sections describe the approach to finding frequent itemsets and association rules in blood count data from COVID-19 cases. The dataset used is described in section II. An overview of the methodology, the algorithm used, and steps taken is outlined in section III. The section Results IV highlights and discusses the obtained results using different measures. Lastly, a summary is given, and conclusions are drawn in section V.

II. DATA DESCRIPTION

The dataset was obtained from Kaggle and was published under the headline "Diagnosis of COVID-19 and its clinical spectrum" with the aim to support clinical decisions based on blood count data in the current COVID-19 pandemic [7]. The clinical data were collected and made available by the Hospital Israelita Albert Einstein in Sao Paulo, Brazil. A study using these data for Topological Data Analysis was published as preprint by Dlotko and Rudkin from Swansea University, UK [8].

Table 1 gives a sample description of the original dataset which contained 5644 patient samples (rows) and 111 features (columns). Due to space limitations, only a couple of features are shown in Table 1 below.

TABLE 1. DESCRIPTION OF DATASET

Attribute	Type	Example Value	Description
PATIENT_ID	Object	44477f75e8169d2	Record identifier
PATIENT_AGE_QUANTILE	Integer	13	Quantile dividing age into bins
SARS_COV_2_EXAM RESULT	Binary variable	positive or negative	Outcome of SARS-CoV-2 test
PATIENT ADMITTED TO REGULAR WARD	Integer	1=yes, 0=no	Patient admitted to regular ward
PATIENT ADMITTED TO SEMI-INTENSIVE UNIT	Integer	1=yes, 0=no	Patient admitted to semi-intensive unit
PATIENT ADMITTED TO INTENSIVE CARE UNIT	Integer	1=yes, 0=no	Patient admitted to intensive care unit

For the exploratory data analysis, the data were loaded into the Jupyter notebook [9] environment and explored using Python 3 programming software [10] and its library pandas [11]. The visual programming software Orange [14] was also used for data exploration, visualization and subsequent rule-based data mining.

The dataset contained normalized measurements from patients' blood samples including blood parameters like hematocrit, hemoglobin and the counts of different blood cells with a mean of zero and a unit standard deviation. Further, the testing for the presence of other pathogens such as influenza and other respiratory viruses was recorded. The exploratory data analysis yielded 105 columns with missing values. Many features had an extensive amount of values missing with more than 90% und up to 100%. With the hospital being a stressful environment, especially during the peak of the pandemic, it is understandable that these data were sparse. Feature columns with more than 90% of missing data were deleted from the dataframe using pandas operations in IPython. Also, columns with data on the test results of other viruses were dropped because the focus of this study were blood parameters. For subsequent data mining analysis, the remaining 19 features were used.

Investigating the SARS-CoV-2 test results, showed that 5086 patients in the dataset were tested negative whereas 558 patients had a positive result for the virus. From the people admitted to hospital, 79 were admitted to the regular ward, 50 to the semi intensive care unit and 41 to the intensive care unit (see Table 2). The rest of the positive tested people were presumably treated at home in self-quarantine.

Table 2. SARS-CoV-2 Test Results and Hospital Admittance for Patients in Blood Count Data

Negative SARS-CoV-2 Test Result	**Positive SARS-CoV-2 Test Result**		
5086	558		
	Regular ward	***Semi ICU***	***ICU***
	79	50	41

Age is a critical parameter for the outcome of a COVID-19 disease with a higher death rate in older people [15]. In order to deal with the parameter age, the age has been recorded within a certain range. There were 19 age quantiles with a bin width of 5 years each, e.g. quantile 1 included patients from 0 to 5 years of age. The distribution of the feature patient age can be seen in Fig. 1, with age quantile 19 being the most frequent one. The average age of patients with a negative SARS-CoV-2 test result was in quantile 9.17 whereas the average age of patients with a positive SARS-CoV-2 test result was in quantile 10.63.

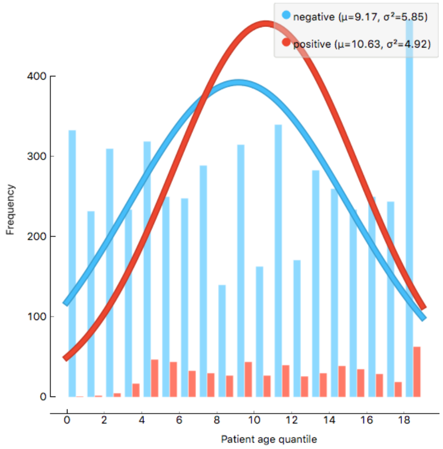

Fig. 1. Age distribution of patients labelled according to negative and positive SARS-CoV-2 test result.

The dataset further contained information on various blood characteristics such as hemoglobin, the protein that transports oxygen within red blood cells [16], or counts on different blood cells. All descriptors varied within a certain range that was negative-bounded on one side and in the low positive range on the other side due to the normalization process. Using descriptive statistics, the feature columns were explored for their distribution in means, standard deviation, minimum and maximum values and different percentile measures. Table 3 shows these statistical values for a selected subset of features.

TABLE 3. DESCRIPTIVE STATISTICS ON FEATURE COLUMNS (EXAMPLE)

	Patient Age Quantile	Hematocrit	Hemoglobin	Platelets
count	5644.000000	6.030000e+02	6.030000e+02	6.020000e+02
mean	9.318391	-2.186214e-09	-1.601319e-08	-3.535003e-10
std	5.777903	1.000830e+00	1.000830e+00	1.000832e+00
min	0.000000	-4.501420e+00	-4.345603e+00	-2.552426e+00
25%	4.000000	-5.188074e-01	-5.862439e-01	-6.053457e-01
50%	9.000000	5.340703e-02	4.031596e-02	-1.217160e-01
75%	14.000000	7.171751e-01	7.295320e-01	5.314981e-01
max	19.000000	2.662704e+00	2.671868e+00	9.532034e+00

III. METHODOLOGY

In order to find relationships between individual items, the open-source machine learning and data mining software Orange was used. It provides a visual programming approach to data mining [17] and was preferred over a IPython due to the lack of knowledge of adequate rule mining libraries. Finding association rules between data instances as well as frequent item sets can be accomplished with the Associate add-on in Orange [18]. Association rule mining is typically applied in market basket analysis [6] but has been used in other research areas where relationships between variables in a large dataset are of interest such as in the field of bioinformatics [19], disease diagnosis and text mining. The baskets in this study are not transactions, but individual patients. The items are various measures of blood markers such as the count of different blood cells (e.g. platelets, neutrophils, basophils, etc.). The goal of this study is to find association rules between blood parameters of COVID-19 patients that have user-specified minimum support and minimum confidence and thus are related. In order to find associations between blood parameters leading to severe COVID-19 illness, the dataset was filtered for patients with SARS-CoV-2 positive results using Python and its pandas library.

The stepwise methodology taken in this study is visualized in Fig. 2. The data was subjected to a preprocessing step involving the discretization of continuous variables. Association rules were found using the Orange Association Rules module. This widget implements the FP-growth [21] frequent pattern mining algorithm with bucketing optimization [22]. For generating rules, initially a maximum confidence of 100 and a minimum support of 10 were chosen following the increasing of the support value und the decreasing of the confidence value. Frequent itemsets were found using the Frequent Itemset module of Orange. The results will be described in section IV.

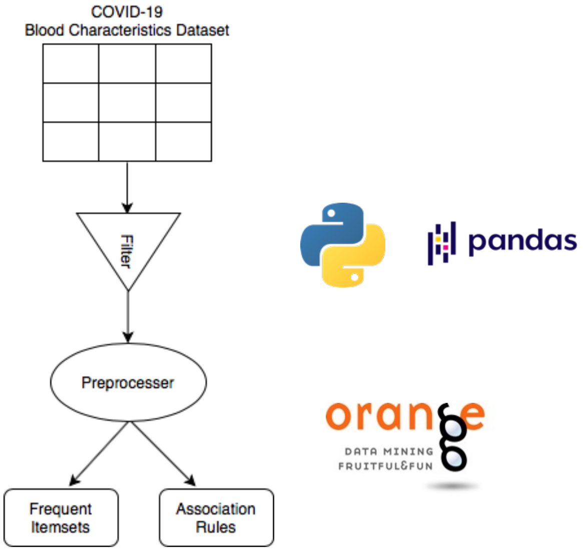

Fig. 2. Flowchart describing the approach of association rule mining and the used software.

IV. Results and Discussion

This section documents and explains the results of the frequent itemsets (section B) discovered and association rules generated (section A).

A. Frequent Itemsets

In order to find collections of items that show up together frequently, the minimum support was varied.

Setting the minimal support (minsup) to 10% and the minimum number of items that occur together to 5, the algorithm finds four itemsets (see Table 4 for an example). Investigating the individual items in each of the four itemsets showed that the items are almost the same. The items differ in their age quantile ranges, but are consistent in a positive SARS-CoV-2 test result and no admittance to neither regular ward, semi-intensive unit or intensive care. For patients in the age quantile below 6.5, as shown in Table 4, this makes sense because in most cases elderly people are admitted to hospital and younger ones can recover from the disease at home. However, the itemsets found are not satisfactory because further itemsets contained higher patient age quantiles together with a positive SARS-CoV-2 test and no hospital admittance at the same support level. This makes it suspicious because some older patients might need more intense care.

TABLE 4. FREQUENT ITEMSETS WITH MINSUP=10%

	Items	Support	%
ITEMSET	Patient age quantile =< 6.5	148	26.52
	SARS-CoV-2 exam result = positive	148	26.52
	Patient admitted to regular ward (1=yes,0=no)=0	148	26.52
	Patient admitted to semi-intensive unit (1=yes,0=no)=0	148	26.52
	Patient admitted to intensive care unit (1=yes,0=no)=0	147	26.34

B. Finding association rules

In order to find rules between frequent items that logically imply, the minimal support was varied starting from 0.1% and minimum confidence (minconf) to 90%. The strength of the rules is given in support, confidence, coverage, strength, lift and leverage (Table 5). The aim of this study was to find blood parameters as antecedents that imply intensive care for SARS-CoV-2 positive patients. However, no rules indicating this relationship were found. Most rules were of the following form

Patient admitted to semi-intensive unit (1=yes, 0=no)=0 → SARS-Cov-2 exam result=positive

and had a support of 98%. While this doesn´t yield novel insights from a clinical perspective it makes sense based on the data available. Most of the attributes used had complete information on hospital care and the SARS-CoV-2 test result.

The red blood cell count and the range of basophils have been the only blood parameters to become part of a rule and indicative of a positive virus test or no admittance to semi-ICU (Table 5). However, the support for these rules is quite low (0.2%).

TABLE 5. ASSOCIATION RULES WITH 0.1% MINSUP AND 90% MINCONF

Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		Consequent
0.002	1.000	0.002	558.000	1.000	0.000	Patient age quantile=< 6.5, Red blood Cells=0.295938 - 0.948261, Basophils=≥ -0.071037	→	SARS-CoV-2 exam result = positive
0.002	1.000	0.002	522.000	1.069	0.000	Patient age quantile=< 6.5, Red blood Cells=0.295938 - 0.948261, Basophils=≥ -0.071037	→	Patient admitted to semi-intensive unit (1=yes, 0=no)=0

V. CONCLUSIONS

This study provides an attempt to apply association rule mining to blood count values of COVID-19 patients in order to infer whether they would need intensive care for further treatment. Using the FP-Growth algorithm of the Orange Associate Module, many rules and frequent itemsets were generated. However, none of these rules were indicative of which blood parameters were reliable prognostic markers to make the association of the requirement of hospital care. The reason for these results might be the dataset which only sparsely had the values for the blood parameters recorded. Most data entries were concerned with the categorical labels of admittance to different care regimes (regular ward, semi-ICU and ICU) or labeled with positive or negative for the test result. Thus, these feature items were over-represented in the dataset and hindered the procedure of finding meaningful rules. Thus, for future analysis, the data collection process needs to be optimized and the rule-based analysis redone using a more complete dataset.

REFERENCES

[1] "Coronavirus." https://www.who.int/emergencies/diseases/novel-coronavirus-2019 (accessed Jun. 11, 2020).

[2] D. S. Hui *et al.*, "The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health — The latest 2019 novel coronavirus outbreak in Wuhan, China," *Int. J. Infect. Dis.*, vol. 91, pp. 264–266, Feb. 2020, doi: 10.1016/j.ijid.2020.01.009.

[3] "Q&A on coronaviruses (COVID-19)." https://www.who.int/news-room/q-a-detail/q-a-coronaviruses (accessed Jun. 11, 2020).

[4] CDC, "Coronavirus Disease 2019 (COVID-19) – Symptoms," *Centers for Disease Control and Prevention*, May 13, 2020. https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html (accessed Jun. 11, 2020).

[5] CDC, "Coronavirus Disease 2019 (COVID-19) - Transmission," *Centers for Disease Control and Prevention*, Jun. 01, 2020. https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/how-covid-spreads.html (accessed Jun. 11, 2020).

[6] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," 1993, pp. 207–216.

[7] "Diagnosis of COVID-19 and its clinical spectrum." https://kaggle.com/dataset/e626783d4672f182e7870b1bbe75fae66bdfb232289da0a61f08c2ceb01cab01 (accessed Jun. 11, 2020).

[8] P. Dlotko and S. Rudkin, "Covid-19 clinical data analysis using Ball Mapper," Intensive Care and Critical Care Medicine, preprint, Apr. 2020. doi: 10.1101/2020.04.10.20061374.

[9] F. Perez and B. E. Granger, "IPython: A System for Interactive Scientific Computing," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 21–29, 2007, doi: 10.1109/MCSE.2007.53.

[10] Pilgrim, M., & Willison, S. (, "Dive Into Python 3," vol. Springer, .

[11] W. McKinney, "Data Structures for Statistical Computing in Python," presented at the Python in Science Conference, Austin, Texas, 2010, pp. 56–61, doi: 10.25080/Majora-92bf1922-00a.

[12] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, May 2007, doi: 10.1109/MCSE.2007.55.

[13] Michael Waskom *et al.*, *mwaskom/seaborn: v0.8.1 (September 2017)*. Zenodo, 2017.

[14] J. Demšar and B. Zupan, "ORANGE: DATA MINING FRUITFUL AND FUN," p. 4.

[15] E. Mahase, "Covid-19: death rate is 0.66% and increases with age, study estimates," *BMJ*, vol. 369, Apr. 2020, doi: 10.1136/bmj.m1327.

[16] "Hemoglobin," *Wikipedia*. Jun. 09, 2020, Accessed: Jun. 13, 2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Hemoglobin&oldid=961650993.

[17] J. Demsˇar *et al.*, "Orange: Data Mining Toolbox in Python," p. 5.

[18] "Orange Data Mining - Association Rules in Orange." https://orange.biolab.si/blog/2016/04/25/association-rules-in-orange/ (accessed Jun. 12, 2020).

[19] S. Naulaerts *et al.*, "A primer to frequent itemset mining for bioinformatics," *Brief. Bioinform.*, vol. 16, no. 2, pp. 216–231, Mar. 2015, doi: 10.1093/bib/bbt074.

[20] R. Agrawal, "Fast Algorithms for Mining Association Rules," p. 13.

[21] "Mining frequent patterns without candidate generation | Proceedings of the 2000 ACM SIGMOD international conference on Management of data." https://dl.acm.org/doi/10.1145/342009.335372 (accessed Jun. 13, 2020).

[22] R. C. Agarwal, C. C. Aggarwal, and V. V. V. Prasad, "Depth first generation of long patterns," in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '00*, Boston, Massachusetts, United States, 2000, pp. 108–118, doi: 10.1145/347090.347114.