Technical Communication

To: Emily N Smith

From: Yeonhak Kim

Subject: Technical Analysis of the Research Paper - COVID-19 Symptoms Detection Using ID3 Algorithm

Date: January 26th, 2022

The main objective of the memo is to deliver technical analysis over the selected research paper that proposes machine learning model for predicting COVID-19 positive cases by implementing ID3 algorithm.

While the paper successfully presents the overall concept and structure of the experiment conducted, it fails to meet the level of details expected to see in any research paper in the perspective of its experiment and observation. Throughout the memo I will elaborate how the research paper addressed the subject of matter in a proficient way and which part of it shows lack of details.

Analysis

First, the title of the research paper is "COVID-19 Symptoms Detection Using Iterative Dichotomiser (ID3) Algorithm: A Greedy Method" by four authors whom the main authors published the paper are Hassan Adamu and Azman Samsudin. Publication date of April 5th, 2021.

As the main objective of the research paper is to deliver its experimental observation and results to the world, the targeted audience seems to range from university students to following scholars and industry researchers. The paper has a particularly good abstract that summarizes the overall content in a concise and clear manner as well as proposing its motivation behind the experiment. As the audience is mostly targeted to the entities related to the research fields, the format and general structure of the paper are very formal and understandable.

The structure of the document is divided into 8 parts: Introduction, Brief explanation of COVID-19 virus, Symptoms ID3 algorithm, Experiment results, Experiment observations, Conclusion and References.

1. Introduction

- It briefly explains what the intent of the paper is and the background information about the main algorithm being proposed which helps audiences to understand what the ID3 is about.

The paper elaborates how greedy algorithms can solve a complex combinatorial problem and can be applied to the COVID-19 screening process such that it can help ease the load of testing procedures.

2. Background of COVID-19

- This section covers the background of COVID-19 virus. The section would be especially useful to the audiences in 5-10+ years in the future who needs to be explained about what the virus is and how it

affects the human body. The paper also compares with other predecessor viruses such as SARS and MERS which can be also helpful in understanding the virus.

Having a figure in this section was helpful in visualizing the virus.

However, this section could have been more dedicated to explaining in-depth symptoms and the explicit characteristics of the COVID-19 virus instead of briefly covering different viruses and their historical aspects.

3. Symptomatic Diagnosis of COVID-19 via CT

- Many of the COVID-19 symptoms and effects are elaborated in this part. How the testing procedure works, and which information was collected from the patients. It also mentions getting information and classifying the result via CT which I personally think is unnecessary part of this research paper. Overall, this section could have been added to the previous section excluding the CT parts which is excess information here.

4. ID3 Algorithm

ID3 is a greedy version of the decision tree model. The model splits the class node in order of highest information gain to the lowest information gain. The term greedy comes from the model attempts to pick the entity that is most beneficial (in terms of prediction) at the given time. Information gain is the term for how much information can be extracted or estimated from the given input values.

The paper clearly explains how the algorithm behaves as well as proposing the implementation of the algorithm (via pseudocode). Two approaches have been used in ID3 implementation which are holdout approach and cross validation approach. Comparing these two methods are the main objective of the experiment and results are discussed in the later sections.

One detail that is crucial and which is missing in this section is the specific example of the sample data. It references the data set from the nCov2019 repository on GitHub but, the figure presented only has a brief description of them. Audiences will need to visit the repository to see how the data set looks like and which features were considered as to be factors that contribute to COVID-19 positive cases.

The part where in table 2 in this section must have been elaborated what symptoms 1 to 5 are.

Moreover, the paper did not specify the size of the data set, how the ratio is in between positive cases and negative cases. This is important because depending on how the distribution of labels is in the given data set, the model can be trained to lean on one side.

The number of negative cases will outnumber the number of positive cases we can expect the model to be leaned on the negative side.

Last to mention in this section is it made a spelling error. "Native Bayes" should be "Naive Bayes".

5. Results

The experiment was conducted in two methods (holdout vs cross-validation) where each method was approached in three separate ways. Each way of approach had differentiated the ratio between the training data set and testing data set.

The diagrams given shows the clear result of each method and it seems to be straight forward in comparing these two results.

6. Observation

The observation for the conducted experiment was concise and comparisons were clear. Having criteria for both accuracy and runtime are good factors to make a judgement between the two methods.

But I would like to add that in this section it should have talked more about the data set and validity of the accuracy for both methods (holdout and cross-validation).

The reason is first, as mentioned earlier, the model is expected to be leaned towards the negative case. Both the training and test data sets will naturally have an overwhelming number of negative cases than the positive cases. Thus, it is highly likely that the model will end up hitting the negative cases and the data being fed into the model will be mostly negative cases. Which means the accuracy will be high because both the model and fed input data set are leaned towards to the negative cases.

To measure the model more accurately, the experiment should also have added a test where only positive cases of the input data set are given into the model.

Another point to be mentioned is that the features(symptoms) selected as data could be misleading. Since the external symptoms of COVID-19 is like that of the flu there would be an edge case where the symptoms overlap to flu and positive cases.

7 & 8. Conclusion and References

The conclusion is simple and concise as well as the formatting for the references.

Recommendation

Although the audience of this paper is meant to be scholars and researchers it seems like the paper is more suitable for entry level of researchers or students who are interested in building new machine learning models.

Surely the paper is not intended for everyone. Unless one has explicit knowledge in this scientific field it will be extremely hard to understand the content.

Overall, the paper was concise and motivational.