Prediction and Analysis of Corona Virus Disease (COVID-19) using Cubist and OneR

R.V.S.Lalitha, J.Divya Lalitha, K.Kavitha, RamaReddy.T, Rayudu Srinivas, Challapalli Sujana

Department of C.S.E, Aditya College of Engineering & Technology, Surampalem
Department of C.S.E, Gokaraju Rangaraju Institute of Engineering & Technology(A), Hyderabad
Department of C.S.E, Aditya Engineering College(A), Surampalem
Department of I.T., Pragati Engineering College(A), Surampalem

{rvslalitha@gmail.com, lalitha0517@gmail.com, kavitha.bits@gmail.com, ramatreddy@gmail.com, srinivas.rayuda@aec.edu.in, sujana.challapalli@gmail.com}

Abstract. Inflammatory syndrome usually occurs with the adults who have already affected with corona virus. Multi system syndrome in children may affect different parts of the body like lungs, liver, kidneys and brain. The level of current infection can be known by proceeding with two kinds of tests, namely viral tests and anti body tests. Viral tests determine the level of current infection and antibody tests are about past infection. Though decision on test will be provided by health care units, there will be an uncertainty in evaluation of results. The variation in potential of antibodies will influence the resulting parameters. The effect of inaccurate test results is to be traced, because people may suffer from reinfeciton in their bodies over a period of time. In this paper the uncertainty analysis is carried out using Cubist and OneR algorithms. In the cubist algorithm leaf nodes can be analyzed using regression models. Accuracy of the predictions are analyzed using contingency table and false positives and true negatives are tracked using confusion matrix. This analysis assists in determining the trans conditional state of the disease with dynamism. With the analysis carried out cubist proves to be the best algorithm.

Keywords: Cubist, OneR, Rule based Learning, Confusion Matrix, Contingency Matrix.

1 Introduction

Corona viruses cause illnesses like cold, respiratory problems such as Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome (MERS). The greatest need of current situation of pandemic diseases is evaluation of test results. In order to make test result vigilant, the analysis of symptoms is to be carried out resourcefully. Even though there are antibodies persisting in the body, there may be a chance of getting reinfected. Since long the attention had been paid towards the people who had developed antibodies for raising immunity levels to the virus. Anti bodies are y shaped proteins that lie in certain regions of the body. Anti bodies are produced by immune cells also known as B cells. They fight against foreign body cells for not getting infected on the specific region that they bind with. There is an imbalance between the virus and the immune cells that fight against virus. Owing to
this, the results will vary. In this paper, the dataset is analyzed using two machine learning algorithms one is Cubist and the other OneRule(OneR). The Cubist algorithm is regression based model and target attribute prediction can be done with several predictor attributes. By submitting the dataset to OneR the impact of individual feature attributes can be well analyzed.

2 Related Work

Jian worked on Random Forest and Cubist algorithms for analyzing shear strength rockfill materials. Comparison of the models were done based on Root Mean Square Error(RMSE), Mean Estimation Error(MAE) and sensitivity analysis for prediction analysis[1,15]. Phan compared Multiple linear regression, Cubist Regression and Random Forest algorithms for estimating temperature of air. The dataset was divided into two parts one is calibration and the other is validation using 70% and 30% ratio respectively. Linear and Multiple linear regression models were used for Ta estimation using MODIS Land Surface Time(LST)[2]. Sathishkumar discussed a rule based regression predictive model for predicting bike sharing based on weather data. The model was evaluated using Cubist, Random Forest, classification and regression trees, KNN and conditional interference tree. Global Positioning System(GPS) enabled mobile application was developed for bicycle renting[3]. 36 machine learning algorithms were compared to predict indoor room temperature of a smart building. Finally, they came to conclusion that ExtraTrees regressor is the best for obtaining highest average accuracy[4,10]. The coronavirus was predicted with regards to accuracy using ARIMA, CUBIST, RIDGE and RF learners. This analysis will assist in decision making systems. Also it was inferred that SVR and stacking ensemble learning are best suitable for forecasting COVID-19 cases[5,11]. Xuan developed highly accurate artificial intelligence model for predicting slope in open pit mines. Author proposed M5Rules-GA model and proved to be the best in comparison with FFA SVR, ANN PSO, ANN ICA, ANN GA and ANN ABC [6,14]. Yaghoub Dabiri proposed a real time left ventricular mechanics simulator for analyzing both passive and active behavior of myocardium during diastole.LV pressures are estimated using XGboost and Cubist. They concluded that Cubist results are good when compared to Gradient Boost algorithm [7]. The transformations of portrait photographs are analyzed using deep learning and image processing techniques. Guanyu Lian worked with online selfie photographs [8,9,12].

3 Methodology

World Health Organization (WHO) emphasized that nobody can give confirmation about the period of protection that antibodies give. Expanding on this point, it is required to track the inaccurate tests by observing the error rates of tests based on the statistical data collected. The uncertainties caused due to inaccurate assumptions can be overcome by analyzing the data with different algorithms. Evidently, the algorithm that yields less error will be considered as solution. The strategic solution derived can
also be reassessed for obtaining better results. The problem is addressed in 2 phases:
1. Applying Cubist algorithm to predict target value with ‘n’ number of input variables that experiments on rules. 2. Using OneR evaluates accuracy of the prediction of each feature in decreasing order.

3.1. Modeling Using Cubist
The dataset is downloaded from kaggle and used for analysis. This contain the following information:
'data.frame': 374 obs. of 9 variables:
$ id : int 1 2 3 4 5 6 7 8 9 10 ...
$ country : chr "China" "China" "China" "China" ...
$ gender : chr "male" "female" "male" "female" ...
$ age : int 66 56 46 60 58 44 37 39 56 ...
$ visitingWuhan: int 1 0 0 1 0 0 1 1 1 ...
$ fromWuhan : int 0 1 1 0 0 1 1 0 0 ...
$ symptom1 : int 0 0 0 0 0 0 0 0 0 ...
$ symptom : chr "" "" "" "" ...
$ death : int 0 0 0 0 0 0 0 0 0 ...

Fig.1. Analysis of symptoms using WEKA.

The frequency count of the symptoms is shown in Fig.1. Cubist algorithm is applied for the dataset chosen, for conditional outcomes. The spread of the disease
may occur with people of age>60 and people visiting or coming from Wuhan as per the dataset. Also the deaths may cause with or without symptoms. Considering these factors, the rules are framed based on the current scenario and analyzed the outputs. The impact of disease attack of corona is analyzed with the rules as mentioned below:

Rule 1: age > 60 & visitingWuhan == 1 || fromWuhan == 1 & death == 1 [Fig.2]
Rule 2: age > 60 & visitingWuhan == 1 || fromWuhan == 1 & death == 0 [Fig.3]
Rule 3: age > 60 & symptom1 == 0 & death == 1 [Fig.4]
Rule 4: age > 60 & symptom1 == 1 & death == 0 [Fig.5]
Using cubist method in R, 299 cases are considered (80% training and 20% test data) with mean 0.1 and range 0 to 1 with estimated error 0.0 for applying one rule. Rule 1: [299 cases, mean 0.1, range 0 to 1, estimated err 0.0] Fig.2-5 represents the rule based analysis based on Cubist algorithm.
3.2. One Rule (OneR)

The implication of disease analysis is computed using OneR for evaluating predictions and is interpreted as below:

*OneR algorithm for resolving uncertainty:*

For each symptom
For each value of the symptom
Count for number of cases becoming positive
Observe the frequent symptoms for positive cases
Create rules based on the observations made
Compute error in each prediction
Select the predictor for which it is less error prone

The following are the observations obtained by implementing OneR algorithm. The OneR evaluation model is depicted in Fig.6. The number of people died off with or without symptoms in Fig.7, and the frequency of death occurred or not in Fig.8.

3.2.1. Contingency table for finding accuracy

Table.1. Contingency Table

| Death      | (0,88) | (88,177) | (177,264) | (264,352) | (352,440) | Sum |
|------------|--------|----------|-----------|-----------|-----------|-----|
| 0          | *63    | *65      | *58       | *72       | *75       | 333 |
| 1          | 24     | 15       | 0         | 0         | 0         | 39  |
| Sum        | 87     | 80       | 58        | 72        | 75        | 372 |

Contingency table is constructed for evaluation of association rules based on the symptoms. The number of deaths occurred in each interval is listed in Table.1.
3.2.2. Confusion matrix for tracking of false positives and true negatives [Table 2.]

Table 2 Confusion matrix

|                   | Actually negative(0) | Actually positive(1) |
|-------------------|-----------------------|----------------------|
| Predicted negative(0) | 335                   | 35                   |
| Predicted positive(1)  | 2                     | 2                    |

The accuracy based on rules using OneR analysis is 89.52 \%(333/372)\), error rate is 0.1043(39/374) and Maximum occurrences in each column is represented by ‘*’ using contingency table. Also the accuracy of prediction is 89.57 \%(335/374) with confusion matrix. The OneR model is explored as given below:

![Fig. 6. OneR model diagnostic plot](image1)

![Fig. 7. Disease clustering based on symptoms](image2)
The death rate analysis is calculated by taking bins, which discretizes the numerical data into clusters. In this scenario, number of bins is taken as 2, stating ‘0’ for alive and ‘1’ for death with people having symptoms. Like that we can pose and analyze rules for any dataset with any number of conditions. Data imbalance arises due to uneven distribution of instances for various classes. Data imbalance problem is analyzed in this paper by forming disease clusters based on symptoms (Fig.1 and Fig.7).

4 Results and Discussion

Based on computational analytics explored above the following observations are inferred and are tabulated in Table.3. as listed below.

Table.3. Comparative analysis of Cubist and OneR

| Parameter       | Cubist                                                                 | OneR                                                                 |
|-----------------|------------------------------------------------------------------------|---------------------------------------------------------------------|
| Purpose         | Regression model by pruning subsets of tree based on rules             | Baseline for classification using which number of deaths is counted using bins. |
| Accuracy        | Estimated error 0, average and relative errors are 0 and correlation coefficient 1. | Error rate: 0.1043 Accuracy: 89.52%                                  |
| Removed instances | 0                                                                       | Two instances are removed due to missing values                       |
| Ordering        | Accuracy will be evaluated based on number of rules                   | Accuracy of the attribute is in decreasing order                      |
| False predictions | 0                                                                       | 2(372 with Contingency table and 374 with confusion matrix           |
| Evaluation      | Evaluation based on training data                                      | Based on total instances                                             |
| Rules           | Only one rule is used. More number of rules can be imposed using ‘rules’ parameter of cubistcontrol method. | 5 rules are framed based on id attribute.                             |
**Conclusion**

With the time varying instances both the algorithms competed with each other in the performance point of view. However, since the accuracy takes highest priority, and the corona disease impact will be based on symptoms prediction analysis should be carried out properly. OneR evaluates accuracy based on the chosen attribute in decreasing order. If the individual feature is more of concern, OneR is preferable and Cubist evaluates large datasets in a very less amount of time. Cubist gives better result as the error rate is zero with reference to OneR algorithm. The analysis carried out so far, helps in facilitating baseline performance analysis of statistical data to the authorities involved in decision making.

**References**

1. Jian Zhou, Enming Li, Haixia Wei, Chuanqi Li, Qiuqiu Qiao and Danial Jashed Armagani, Random Forests and Cubist algorithms for Predicting Shear strengths of Rockfill Materials, Applied sciences,9(8),1621,2019. [https://doi.org/10.3390/app9081621](https://doi.org/10.3390/app9081621)

2. Phan Thanh Noi, Jan Degener and Martin Kappas, Comparision of Linear regression, Cubist regression, and Random Forest algorithms to estimate daily air surface temperature from dynamic combinations of MODIS LST data, Remote Sensing,9(5),398,2017. [https://doi.org/10.3390/rs9050398](https://doi.org/10.3390/rs9050398)

3. Sathishkumar V E, Yongyun Cho, A rule based model for Seoul Bike sharing demand prediction using weather data, European Journal of Remote Sensing, 2020. [https://doi.org/10.1080/22797254.2020.1725789](https://doi.org/10.1080/22797254.2020.1725789)

4. Sadi Alawadi, David Mera, Manuel Fernandez-Deglado, Fahed Alkhabbas, Carl Magnus Olsson & Paul Davidsson, A comparison of machine learning algorithms for forecasting indoor temperature in smart buildings, Energy systems, 2020. [https://doi.org/10.1007/s12667-020-00376-x](https://doi.org/10.1007/s12667-020-00376-x)

5. Matheus Herrique Dal Molin Ribeiro, Ramon Gomes da Silva, Viviana Cocco Mariani, Leandro dos Santos Coelho, Short term forecasting COVID-19 cumulative confirmed cases: perspectives for Brazil, Chaos, Solitons and Fractals, 135,2020,109853. [DOI:10.1016/j.chaos.2020.109853](https://doi.org/10.1016/j.chaos.2020.109853)

6. Xuan Nam Bui, Hoang Nguyen, Yasoon Choi, Trung Nguyen Thoi, Jian Zhou and Jie Dou, Prediction of slope failure in open pit mines using a novel hybrid artificial intelligence model based on decision tree and evolution algorithm, Scientific Reports,10, June 2020, Article number 10,9939. [DOI:https://doi.org/10.1038/s41598-020-66904-y](https://doi.org/10.1038/s41598-020-66904-y)

7. Yaghoub Dabiri, Alex der Velden, Kevin L Sack, Jenny S Choy, Ghassan S Kassab and Julius M Guccione, Prediction of Left Ventricular Mechanics using Machine Learning, Frontiers in Physics, September 2019. [https://doi.org/10.3389/fphy.2019.00117](https://doi.org/10.3389/fphy.2019.00117)
8. Guanyu Lian and Kang Zhang, Transformation of portraits to Picasso’s cubism style, The Visual computer, 36, 799-807, 2020. DOI: https://doi.org/10.1007/s00371-019-01661-2

9. Qidi Peng, Nan Rao, and Ran Zhao, Covariance based dissimilarity measures applied to clustering wide sense stationary ergodic processes, Machine Learning, 108, 2159-2195, 2019. DOI: https://doi.org/10.1007/s10994-019-05818-x

10. Pierre Alquier and Benjamin Guedj, Simpler PAC-Bayesian bounds for hostile data, Machine Learning, 107, 887-902, 2018.

11. Qian, Z., Alaa, A.M. & van der Schaar, M. CPAS: the UK’s national machine learning-based hospital capacity planning system for COVID-19. Mach Learn (2020). https://doi.org/10.1007/s10994-020-05921-4

12. Heinze-Deml, C., Meinshausen, N. Conditional variance penalties and domain shift robustness. Mach Learn (2020). https://doi.org/10.1007/s10994-020-05924-1

13. Raghavendran, C. V., Satish, G. N., Krishna, V. and Basha, S. M. (2020). Predicting Rise and Spread of COVID-19 Epidemic using Time Series Forecasting Models in Machine Learning. International Journal on Emerging Technologies, 11(4): 56–61.

14. K. Kavitha, S. Laxman Kumar, Pudu Pravalika, Kotte Sruthi, R.V.S. Lalitha, N.V. Krishna Rao, Fashion compatibility using convolutional neural networks, Materials Today: Proceedings, 2020, ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2020.09.365.

15. Dr.R.V.S.Lalitha, Dr.G.Jaya Suma, Analyzing the impact of Corona Symptoms using Correlation, Journal of Critical Reviews, Volume 7, Issue 18, pp 281-287, ISSN 2394-5125, June 2020