EasyChair Preprint

№ 3223

Artificial Intelligence Approach to Predict the COVID-19 Patient's Recovery

Aboul Ella Hassanien, Aya Salam and Ashraf Darwish

April 22, 2020

Artificial Intelligence Approach to Predict the COVID-19 Patient's Recovery

Aboul Ella Hassanien[1,2], Aya Salama[2] and Ashraf Darwsih[3,2]

[1]Faculty of Computers and Artificial Intelligence, Cairo University, Egypt
[2]Scientific Research Group in Egypt www.egyptscience.net
[3]Faculty of Science, Helwan University, Egypt

Abstract: Coronaviruse is the new pandemic hitting all over the world. Patients all over the world are facing different symptoms. Most of the patients with severe symptoms die specially the elderly. In this paper, we test three machine learning techniques to predict the patient's recovery. Support vector machine was tested on the given data with mean absolute error of 0.2155. The Epidemiological data set was prepared by researchers from many health reports of real time cases to represent the different attributes that contribute as the main factors for recovery prediction. A deep analysis with other machine learning algorithms including artificial neural networks and regression model were test and compared with the SVM results. We conclude that most of the patients who couldn't recover had fever, cough, general fatigue and most probably malaise. Besides, most of the patients who died live in Wuhan in china or visited Wuhan, France, Italy or Iran.

Key Words: Corona, Multi-layer perception, Support vector machine, Regression, Artificial intelligence.

1. Introduction

Coronaviruse or what so called COVID-19 is the most booming topic now due the enormous spread in a very short duration, causing significant number of deaths [1]. Corona is infectious decease created by severe acute respiratory syndrome coronavirus 2. This virus first appeared in china and then it spread to the whole world killing people everywhere. Most of the patients cannot overcome this decease. Hospitals do not accept all cases due to the limited number of beds specially in the developing countries like Egypt. However, hospitals admit cases based on its severity. Corona is now the common enemy to the whole world [2]. People in the front line need all the possible ways of technology that could help. Artificial intelligence and machine learning already proved their significance many of similar decease. Even now with corona, machine learning and technology in general is one of the main weapons enabling us to face that pandemic virus [3]. Artificial Intelligence has been used in [4,10] to detect where the possible next outbreak is. However, in this work, we are trying to conduct more helpful tool for the white army in

frontlines to be able to predict severity of the cases by estimating the recovery possibility of patients. We trained out models on the data set given by online [5].

The data set given by BoXu et al. showed that in some cases, patients with no symptoms at all died by the Corona virus. For this reason, in this paper we provide preliminary experiments to predict the severity of the case as it may be manipulating to just depend on the obvious symptoms. The number of patients increases exponentially daily. However, symptoms of most of the patients were not recorded in the Data set. Moreover, some records had symptoms but the final state of the patients of weather they recovered or died were not submitted yet. For this reason, we worked only on 108 records with symptoms. In addition to symptoms, age was also given. Moreover, we found that factor of travelling to one of the cities where the virus is spread is also an important factor. Therefore, we added this factor to the inputs to the implemented classifier. These cities are China, France and Italy.

2. Machine Learning: an overview

Machine learning is a main branch of AI. Artificial Neural Network (ANN) and Support Vector Machine (SVM) are the most known machine learning approaches used in multiple domains. ANN simulates the neurons in human brains, where nodes in neural network represent the neurons. For non-linearly separable problems, more than one hidden layers are used and each layer consists of one or more nodes [6]. ASVM are very similar to ANN. However, AVM depends on the kernel trick in which the features are converted into higher dimensional space to be linearly separable. Then, optimization techniques are used to maximize the hyperplane between classes to obtain the best classification results [7].

2.1 Artificial Neural Networks (ANNs)

ANNs are inspired from the biological behavior of brain networks. It has obtained an interest and used in various applications such as pattern and speech recognition and disease diagnosis. The backbone of the ANNs development are the neural network model. Recently, ANNs are become well known and helpful model in some approaches such classification, clustering, prediction and in many disciplines.

ANN works are based on layers of hidden nodes as mentioned before. ANN learns by training on labeled data sets. This way of learning is called supervised learning. Training usually takes several iterations. In each iteration, classification error is computed. This error is used to update weights on outputs nodes. Then these errors are back propagated to update the weights on hidden nodes using chain of first derivatives. This process takes place until minimum classification error is reached or maximum number of iterations [8]. The data analysis factors explain the importance of ANNs which is efficient and successful in providing a high level of capability to be used with complex and non-complex problems. ANNs have an advantage is that it can make models more accurate and easy to use from complex problems. The ANN is a good model that can be used with the medical applications.

2.2 Support Vector Machine (SVM)

SVM is a computational power kernel-based tool for data regression and classification. Compared with other machine learning techniques, SVM has better generalization performance. Therefore, SVM has achieved high level of performance in many real-world applications such as image processing of medical applications.

SVM can be considered as ANN if sigmoid function as activation functions to update weights. However, instead of using multiple layers to solve nonlinear problem, SVM tends to move the features to higher dimensions that it can be separable by hyperplane. Then the goal is to maximize this hyperplane because the more features are separated, the more accurate classification become [9]. Equations from 1 express the SVM equations.

Consider that there are a series of data points $D = [x_i, d_i]_i^n$ where n is the size of data and d_i represents the target value and x_i represents the input space vector of the sample. The SVM estimates the function as given in the following two equations:

$$f(x) = \omega\, \varphi(x) + b \tag{1}$$

$$R_{SVMs}(C) = \frac{1}{2}\omega^2 + C(\frac{1}{n})\sum_{i=1}^{n} L(x_i, d_i) \tag{2}$$

where $\varphi(x)$ is the high-dimensional space feature, b is a scalar, ω is a normal vector and $C(\frac{1}{n})\sum_{i=1}^{n} L(x_i, d_i)$ signifies the empirical error. *b* and *w* can be assessed by equation (2).

3. Data sets characteristics and analysis

The data set from the health report all over the world gave the following information: ID, age, sex, city, province, country, wuhan-0, not-wuhan-1, latitude, longitude, geo-resolution, date-onset-symptoms, date-admission-hospital, date-confirmation, symptoms, lives-in-Wuhan, travel-history-dates, travel-history-location, reported-market-exposure, additional-information, chronic-disease-binary, chronic-disease, source sequence-available, outcome, date-death-or-discharge, notes-for-discussion location, travel-history-binary. Sample of the available data online is shown in table 1.

Table 1. The collected data from hospital reports

ID	age	sex	city	province	country	wuhan 0 _	latitude	longitude	geo_resol	date_onse	date_adm
				Tucuman	Argentina		-26.94	-65.34	admin1		
				Buenos Ai	Argentina		-34.6033	-58.3817	admin1		
				Buenos Ai	Argentina		-34.6033	-58.3817	admin1		
				Buenos Ai	Argentina		-34.6033	-58.3817	admin1		
				Cordoba	Argentina		-31.4167	-64.1833	admin1		
				Cordoba	Argentina		-31.4167	-64.1833	admin1		
				Cordoba	Argentina		-31.4167	-64.1833	admin1		
				Cordoba	Argentina		-31.4167	-64.1833	admin1		
				Cordoba	Argentina		-31.4167	-64.1833	admin1		
				Cordoba	Argentina		-31.4167	-64.1833	admin1		

After processing the data, we found that most of the patients who couldn't recover had fever, cough, general fatigue and most probably malaise. In addition, most of the patients who died live in Wuhan in china or visited Wuhan, France, Italy or Iran. For this reason, we used this data in classifying the patients. In addition, patients who couldn't recover from the Corona virus are above 50. Therefore after processing the data, the attributes that contributed in classification are the following: age, lives in china (Boolean), visited recently Italy (Boolean), china, France or Itan (Boolean), has fever (Boolean) has cough (Boolean), has sore throat (Boolean), has diarrhea (Boolean), has general weakness (Boolean), has nosal problem (Boolean), has headache (Boolean), has malaise (Boolean), has penomedia (Boolean). Sample of the processed data is shown in table 2.

Table 2. The processed data set

throat	pneumon	weakness	sneezing	nosal	diarrhea	breath	head	malaise	traveled	recovered
0	0	0	0	0	0	0	0	1	1	1
1	0	0	0	0	0	0	0	0	1	1
0	0	1	0	0	0	0	0	0	1	1
0	0	0	1	0	0	0	0	0	1	1
0	1	0	0	0	0	0	0	0	1	1
0	1	0	0	0	0	0	0	0	0	1

Visualization of data is also given in figure 1. This figure represents histogram for each attribute. As obvious in this chart that patients with age above 50 are more vulnerable to get infected. It also shows that most of the studied cases didn’t suffer from sneezing o diarrhea or nosal

problems. However, the more common symptoms are fever and cough.

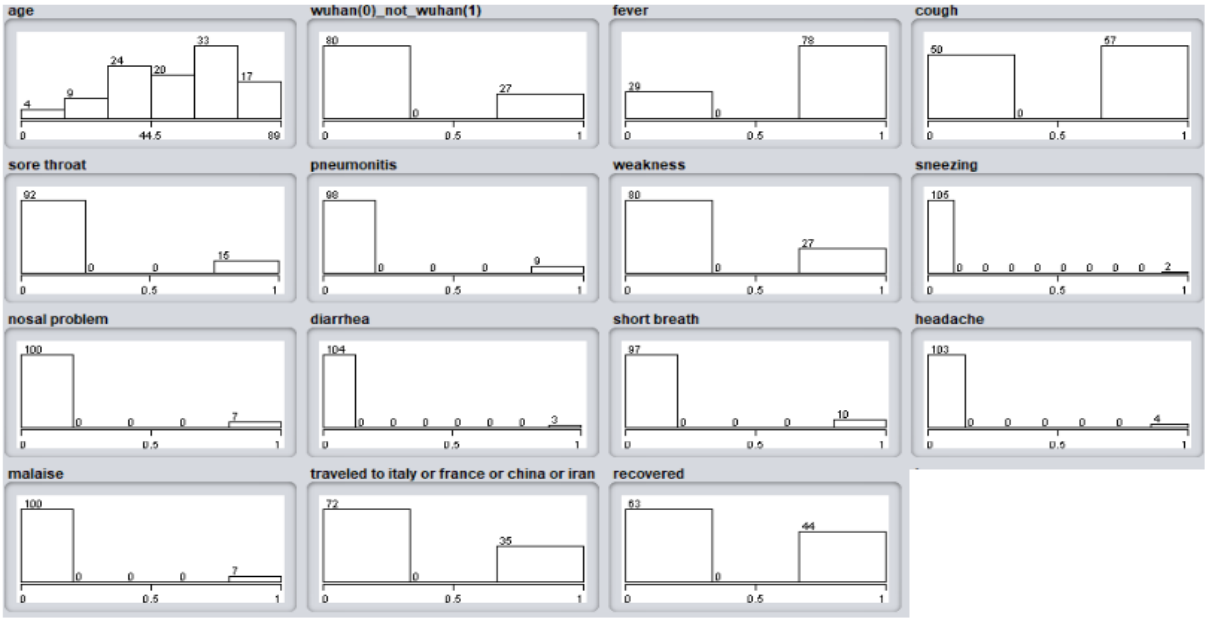

Fig. 1. Data visualization

All the data set given by WHO organization or any other responsible committee concerned about patients diagnosed positively with Corona only. For this reason, we could not use this data for deciding if patients actually have Corona or not.

4. Methodology and results

In this work, we tested artificial neural network, support vector machine and linear regression model in recovery estimation of Corona patients as illustrated in figure 2.

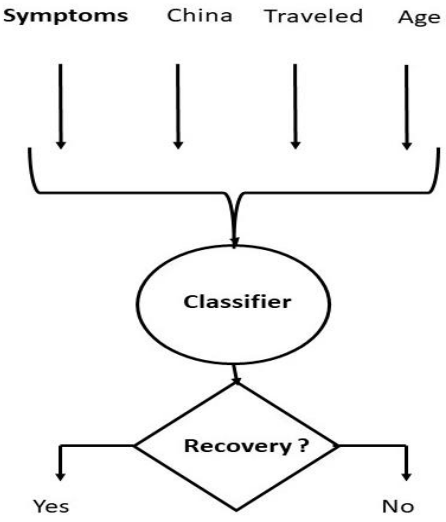

Fig. 2. Summarized chart of the proposed approach-

For the multi-layer perception, learning rate was 0.3, momentum was 0.2, and sigmoid function was used with 10-fold cross validation. All these experiments were carried out using the Weka classifier.

The mean square error and the absolute error of the different conducted experiments are elaborated in table 3.

Table 3. Experimental results

	Mean absolute error	**Root mean squared error**
SVM	**0.21**	0.46
ANN	0.53	0.702
Regression	0.3001	**0.3894**

As obvious, SVM resulted in the minimum absolute error. However, Linear regression resulted in the minimum root mean squared error which is slightly less that resulted from the SVM. For this reason, SVM can be considered as the best model that can be used for recovery prediction. On the other side, Multi-layer perception was dramatically below the other methodologies tested for prediction.

Since most of the patients now are not from China. These experiments were also carried out with removing the Boolean attribute of living in China. As elaborated in table 4, SVM still results in the minimum classification error among all the other classifier. However, ANN performed better after removing the Boolean attribute of living in China.

Table 4. Results after removing China attribute

	Mean absolute error	**Root mean squared error**
SVM	**0.21**	0.46
ANN	0.46	0.61
Regression	0.34	0.41

The experiments were also conducted on the symptoms only so as to cover records without travelling history. However, as illustrated in table 5, there was a significant increase in the classification error as shown in table 5. Therefore, symptoms alone are not enough to decide the severity of the Corona cases.

Table 5. Results after working on symptoms only

	Mean absolute error	**Root mean squared error**
SVM	**0.38**	0.48
ANN	0.48	0.61
Regression	0.36	0.43

Graphical representations of the impact of using different combinations of attributes on classification error are given in figures 3 and 4.

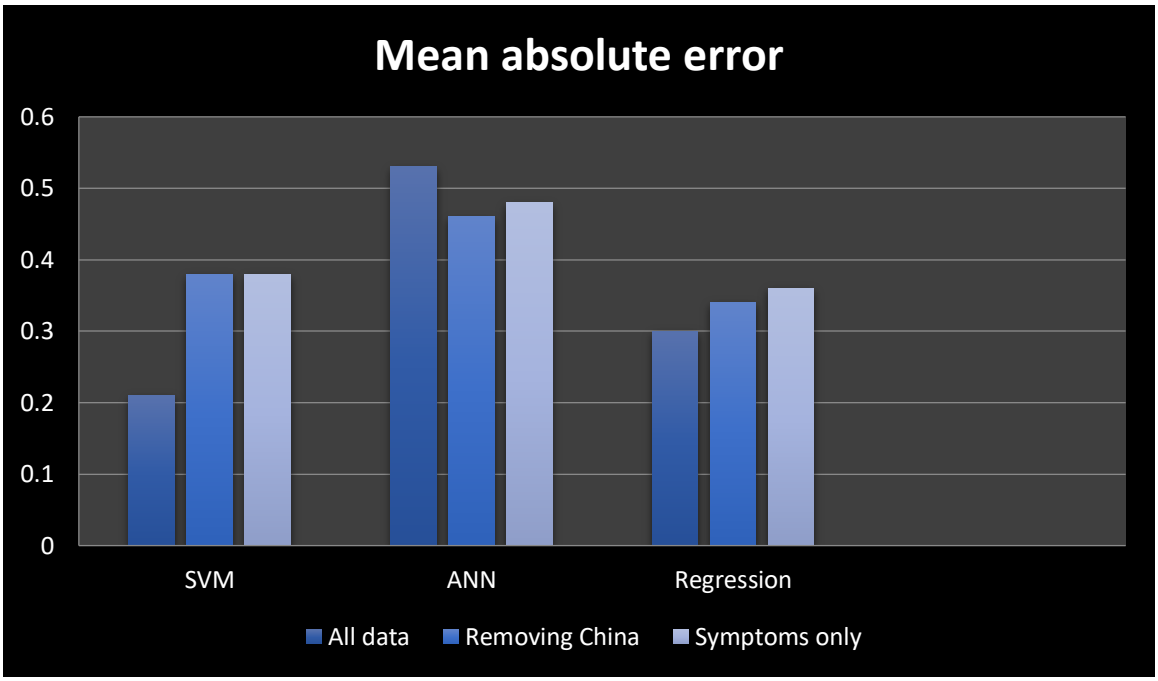

Fig. 3. Graphical representation of the impact of different attributes set on mean absolute error

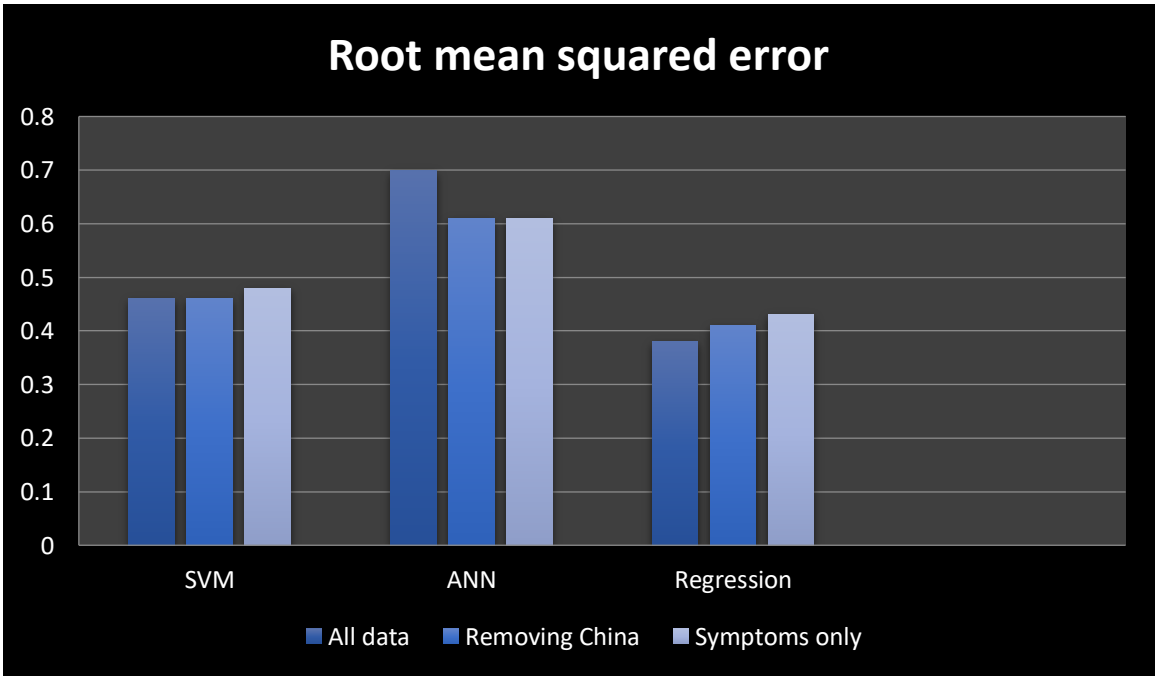

Fig. 4. Graphical representation of the impact of different attributes set on root mean squared error

As obvious in the correlation coefficient matrix given in figure 54. Data are weakly correlated. So we cannot say that if patient has fever, then he should have cough for example.

	age	China	fever	cough	throat	pneumon	weak	sneez	nosal	diarrhea	short brea	heada	malais	travelec
age	1													
China	-0.2	1												
fever	-0.1	-0.08	1											
cough	0.01	-0.23	0.27	1										
throat	-0.2	-0.17	-0.06	0.05	1									
pneun	0.15	-0.02	-0.42	-0.26	0.072	1								
weekr	0.11	0.059	-0.13	-0.06	-0.11	-0.09853	1							
sneez	-0.1	-0.08	0.08	0.13	0.143	-0.04182	0.08	1						
nosal	-0.1	-0.07	0.08	0.02	0.22	-0.08018	0.02	0.24	1					
diarrh	-0.1	0.032	-0.02	-0.07	-0.07	-0.05147	-0.1	-0.02	-0.04	1				
breath	0.17	0.035	-0.09	-0.09	-0.04	-0.0973	0.04	-0.04	0.04	-0.0545	1			
heada	-0.1	-0.11	0.12	0.09	0.204	-0.05972	-0	-0.03	0.15	-0.0335	-0.06327	1		
malaia	-0.1	-0.07	0.16	0.17	0.111	-0.08018	-0.2	-0.04	-0.07	-0.0449	-0.08495	0.15	1	
travel	-0.4	0.191	0.11	0.01	0.235	0.147567	-0.1	0.05	0.14	-0.1184	-0.08699	0.07	-0.02	1
recove	-0.5	0.214	0.04	-0.09	0.21	0.157327	-0.1	0.02	0.09	-0.0269	-0.07257	-0.06	0.009	0.5509

Fig. 5. Correlation coefficient matrixes

5. Conclusion and future work

In this work, different classification models were tested to make the best use of the clinical data provided online to be able to predict the severity of the corona cases. SVM on 15 attributes of symptoms and other relevant information of patient achieved minimum classification error of 0.21 which proves the feasibility of the proposed approach. We also proved that symptoms alone cannot help in deciding the severity of the cases. For future work, if data sets can be gathered by researchers or WHO organization or based on personal efforts to include symptoms and other information of suspects of Corona to be able to diagnose that new corona virus. Moreover, data is changing and is added every minute. As a result, more records can be usable by our model.

References

[1] Bogoch, I. I. et al. Pneumonia of Unknown Etiology in Wuhan, China: Potential for International Spread Via Commercial Air Travel. J. Travel Med. https://doi.org/10.1093/jtm/taaa008 (2020).

[2] Bo Xu., et.al "Epidemiological data from the COVID-19 outbreak, real-time case information, Scientific Data, Nature, Vol. 7, 2020

[3] Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect. Dis. 3099, 19–20 (2020).

[4] Brownstein, J. S., Freifeld, C. & Madof, L. C. Digital Disease Detection - Harnessing the Web for Public Health Surveillance. N. Engl. J. Med. 360, 2153–2157 (2009).

[5] "Epidemiological data from the COVID-19 outbreak, real-time case information", BoXu, BernardoGutierrez, Sumiko Mekaru,, Kara Sewalk, LaurenGoodwin, Alyssa Loskill, Emily L.Cohn, Yulin Hswen, SarahC. Hill , Maria M. Cob, Alexander E. Zarebsk, Sabrina Li, Chieh-HsiWu, Erin, Julia D., , KatelynnO'Brien, SamuelV. Scarpino8, John S. Brownstein,, OliverG. Pybus, David M. Pigott, Moritz U.G. Kraemer. Available at: https://www.nature.com/articles/s41597-020-0448-0.pdf.

[6] Kosko, B. 1992. Neural networks and fuzzy systems. Englewood Cliffs, NJ: Prentice Hall.

[7] Bhavsar, Hetal, and Amit Ganatra. 2012. "Variations of Support Vector Machine classification Technique: A survey." International Journal of Advanced Computer Research 2 (6): 230- 236.

[8] Hopfield, J.J. 1982. Neural networks and physical systems with emergent collective computational properties. Proceedings of the National Academy of Sciences of the USA 79, 2554–88.

[9] Dawkins, Paul. 2017. Paul's Online Math Notes. Accessed September 1, 2017. http://tutorial.math.lamar.edu/Classes/CalcII/EqnsOfPlanes.aspx.

[10] Aboul Ella Hassanien, Lamia Nabil Mahdy, Kadry Ali Ezzat, Haytham H. Elmousalami, Hassan Aboul Ella, (2020) Automatic X-ray COVID-19 Lung Image Classification System based on Multi-Level Thresholding and Support Vector Machine. **doi:** https://doi.org/10.1101/2020.03.30.20047787.