JOURNAL OF AMASYA UNIVERSITY THE INSTITUTE OF SCIENCES AND TECHNOLOGY (JAUIST)

Classification of Covid-19 Dataset with Some Machine Learning Methods

**[1]Yavuz ÜNAL, *[2]Muhammet Nuri DUDAK

*[1]Amasya University, Computer Engineering, Amasya and Turkey

*[2]Amasya University, Institue of Science, Amasya and Turkey

[1]ORCID: 0000-0002-3007-679X
[2]ORCID: 0000-0003-2695-8447

Reseach Article

Received: 05.06.2020 Accepted: 30.06.2020

*Corresponding author: yavuz.unal@amasya.edu.tr

Abstract

The covid-19 disease, which emerged in 2019 and affected the world, caused millions of people to be infected and hundreds of thousands of deaths. This disease has brought heavy workload to doctors and healthcare workers. This workload will be mitigated by machine learning and the development of computer-aided diagnostic systems. Any scientific study about this disease will help get rid of this disease as soon as possible. In this study, naive bayes, k-nearest neighbor, support vector machine, decision tree, which are machine learning classification algorithms, have been applied to covid 19 dataset provided via kaggle site. The best classification accuracy is obtained from the algorithm of support vector machines with 100%.

Key Words: Classification,COVID-19, Machine Learning

Özet

2019 yılında ortaya çıkan ve tüm dünyayı etkileyen covid-19 hastalığı milyonlarca insanın enfekte olmasına ve yüzbinlerce insanın ölümüne sebep olmuştur. Bu hastalık doctor ve sağlık çalışamlarına ağır iş yükü getirmiştir. Bu iş yükünün hafifletilmesi makine öğrenmesi ile bilgisayar destekli tanı sistemlerinin geliştirilmesi ile mümkün olacaktır. Bu hastalıkla ilgili yapılacak her türlü bilimsel çalışma bu hastalıktan en kısa sürede kurtulmaya yardımcı olacaktır. Bu çalışmada kaggle sitesi üzerinden temin edilen covid 19 datasetine makine öğrenmesi sınıflandırma algoritmalarından naive bayes, k-nearest neighbor, support vector machine, random forest, decision trees uygulanmıştır. En iyi sınıflandırma doğruluğu %100 ile destek vektör makinalarından algoritmasından elde edilmiştir.

Anahtar Kelimeler: COVID-19, Makine öğrenmesi, Sınıflandırma,

1. Introduction

This disease which emerged in 2019 in Wuhan, China, has spread quite rapidly. This outbreak has led the World Health Organization to call this infection as a global pandemic. COVID -19 enfected to millions and killed hundreds of thousands (Huang, 2020). It is very difficult to control the spread of the disease because it is a disease that can spread very quickly. Typical clinical features of Covid-19 findings are defined as high fever, respiratory symptoms, and decreased white blood cell, although require specific tests for definitive diagnosis (Guan, 2020). The condition poses considerable hardship for professionals in the healthcare. Thanks to the computer-aided diagnosis for this case, the burden on healthcare workers will lessen and the diagnosis process will speed up. Machine learning methods tend to diagnose several illnesses. (McGoogan, 2020).

A dataset containing data on the Covid-19 disease has been used in this research. Using the WEKA program, naïve bayes, support vector machine, decision tree, k-nearest neighbor algorithms were applied to the data present in this dataset. Classification success for these applied algorithms is provided comparatively.

2. Materials and Methods

2.1. Dataset

The dataset used in this research is obtained from the Kaggle site's data set named "COVID-19 Mexico Patient Health Dataset." This dataset consists of 95839 cases which are formed of 19 attributes and recorded by the Mexican government between 15 January 2020 and 3 May 2020 for data on the Covid-19 disease. This dataset is composed of features as shown in Table 1. (Kaggle.com, 2020). These features are the sex of the patient, the type of the disease (Patient type), the intubated state, the state of pneumonia (Pneumonia), the patient's age (Age), the state of being pregnant (Pregnant), the condition of diabetes patient (Diabetes), Chronic obstructive pulmonary disease status (COPD), asthma status (Asthma), Immunosuppression status (Immunosuppression), Hypertension status (Hypertension), Other diseases status (Other_diseases), Cardiovascular status (Cardiovascular), Obesity status (Obesity), Chronic kidney failure status (Chronic_kidneyfailure) smoker status (smoker), Another case status (Another_case), Covid-19 test status (outcome), intensive care status (icu), Death date or status (Death_date)

Table 1. The COVID-19 mexico patient health dataset features

Feature Name	Feature Type	Range	Description
Sex	Numeric	0-1	1-Women, 2- Man
Patient_type	Numeric	1-2	Type 1, Type 2
Intubated	Numeric	1-99	1 = YES 2 = NO 98/97 = NOT APPLICABLE 99 = NOT AVAILABLE
Pneumonia	Numeric	1-99	1 = YES 2 = NO 98/97 = NOT APPLICABLE 99 = NOT AVAILABLE
Age	Numeric	0-113	1 = YES 2 = NO 98/97 = NOT APPLICABLE 99 = NOT AVAILABLE
Pregnant	Numeric	1-98	1 = YES 2 = NO 98/97 = NOT APPLICABLE

Diabetes	Numeric	1-98	1 = YES 2 = NO 98/97 = NOT APPLICABLE
Copd	Numeric	1-98	1 = YES 2 = NO 98/97 = NOT APPLICABLE
Asthma	Numeric	1-98	1 = YES 2 = NO 98/97 = NOT APPLICABLE
Immunosuppression	Numeric	1-98	1 = YES 2 = NO 98/97 = NOT APPLICABLE
Hypertension	Numeric	1-98	1 = YES 2 = NO 98/97 = NOT APPLICABLE
Other_diseases	Numeric	1-98	1 = YES 2 = NO 98/97 = NOT APPLICABLE
Cardiovascular	Numeric	1-98	1 = YES 2 = NO 98/97 = NOT APPLICABLE
Obesity	Numeric	1-98	1 = YES 2 = NO 98/97 = NOT APPLICABLE
Chronic_kidney_failure	Numeric	1-98	1 = YES 2 = NO 98/97 = NOT APPLICABLE
Smoker	Numeric	1-98	1 = YES 2 = NO 98/97 = NOT APPLICABLE
Another_case	Numeric	1-98	1 = YES 2 = NO 98/97 = NOT APPLICABLE
Outcome	Numeric	1-3	1 = COVID-19 POSITIVE 2 = COVID-19 NEGATIVE 3 = NOT APPLICABLE
Icu	Numeric	1-99	1 = YES 2 = NO 98/97 = NOT APPLICABLE 99 = NOT AVAILABLE
Death_date	Numeric	15-01 2020 03-05-2020	9999-99-99 = Live

3. Results and Discussion

Aside from the features in Table-1, we added a column indicating whether the patients were dead or not. This column is a numeric type feature and defined as a class within the arff file of the Weka program.

2.2. Machine Learning Algorithms

This chapter discusses the theoretical infrastructures of machine learning classification algorithms that were employed in this study.

Decision Tree algorithm:

Decision trees are one of the most accurate methods used for classification purposes in data mining. It's often utilized for classification, clustering, prediction models, and to create subgroups within the related research areas of a problem.

The answers to the problem set out in the decision tree are scattered into groups. Decision trees can be produced to make the right decision regarding the cases that have more than one movement.

The decision tree model allows researchers to decide which factors to address during the decision process and how each factor is linked to the decision's various outcomes and the past. The model which has been formed in decision trees is very apparent and concise (Quinlan, 1993).

Random Forest algorithm:

Random Forest is one of the supervised learning algorithms. It is suitable for use in classification and prediction models and is considered one of the basic algorithms. Known methods of machine learning are distinguished by their high predictive accuracy and model interpretability (Suchetana, 2017). This can be applied to lost data and unstable data sets.

It is an algorithm that aims to increase the accuracy rate by producing more than one decision tree at the application stage. Creates the decision forest using the decision trees created (Ho, 1995). The reason the random forest algorithm has a high success rate is its low deflection rate and low correlation of trees. The low deflection rate is achieved by creating fairly large trees.

Naïve bayes algorithm:

In 1812 the mathematician Tomas Bayes laid the foundations for this algorithm. Naive Bayes Algorithm is used for classification throughout disciplines such as data mining, pattern recognition, machine learning. The algorithm is trained with training data impartial of one another and based on the training data, whenever new data arrives, it is estimated which group it belongs to (L. Jiang, 2007).

Support vector machine algorithm:

Support vector machines are effectively employed based on statistical learning theory for solving classification and regression problems. SVM is based on Vladimir Vapnik's and Alexey Chervonenkis' theory. It drew the interest of scientists involved in the area of artificial intelligence, after successful implementations in the 90s.

Transformation is executed in SVMs by the nonlinear movement of the vectors captured from a low-dimensional input space into another high-dimensional space. A kernel that determines this transformation is specified for the system (machine or network) that implements the transformation. During classification, the vectors shifted into high-dimensional space and became linearly separable. Within the separating planes, the vector with the maximum distance to classes is ascertained as the most appropriate linear separator (Hsu C. W., 2002).

K-Nearest neighbor algorithm:

This is a basic classification algorithm which is non-parametric. The algorithm is a lazy learning algorithm. There is no learning process here. It doesn't learn based on the training data. When we want to guess, it looks at the class of the closest k-numbered neighbor in the entire data set.

The K-value is calculated at first. This K value represents the number of elements to check. The neighbor of the data whose class should be determined is analyzed by the nearest K value, and the distance is calculated between them. In the distance calculation, Euclidean distance was used. (Agrawal., 2014).

In this study, research was carried out using Support vector machine, Decision tree, Naive Bayes , K-Nearest neighbor algorithm, Random Forest algorithm, based on the classification of Covid-19 feature vectors. Machine learning algorithms were run using WEKA, and the data set was classified with 10-fold cross-validation. The classifiers utilized were evaluated in terms of their performance.

The confusion matrix was used to calculate the performances of the classifiers. The confusion matrix has been shown in Figure 1.

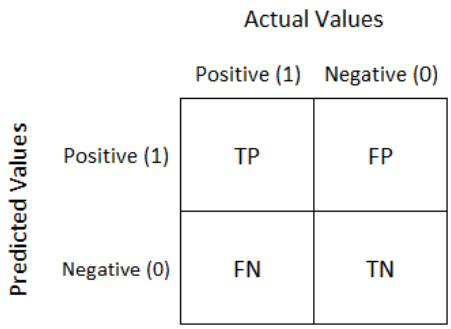

Figure 1. Confusion matrix

TP(True positive) : Observation is positive, and is predicted to be positive.
FN (False Negative): Observation is positive, but is predicted negative.
FP (False Positive): Observation is negative, and is predicted to be negative.
TN (True Negative): Observation is negative, but is predicted positive.
Accuracy, Precision, recall and f-measure formulas has been shown in bellow:

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Precision}=\frac{TP}{TP+FP} \quad (2)$$

$$\text{f-measure}=\frac{2TP}{2TP+FP+FN} \quad (3)$$

$$\text{recall}=\frac{TP}{TP+FN} \quad (4)$$

(Fawcett, 2006) (Powers, 2011)

As a consequence of this analysis, results derived from the SVM classifier are found to be more successful than the results obtained from other classifiers As shown in Table 2 ,BayesNet 99.7704, Naïve Bayes 98.9931, Random Forest 99.9812, J48 99.9906, IBK 99.79 and the support vector machine algorithm provide the best classification accuracy with 100%.

Table 2. Classification results

Classification Tecnique	Accuracy (%)	Precision	F-Measure	Recall	Classfier Name
	99.7704	0,999	0,999	0,999	BayesNet
Naive Bayes	98.9931	0,997	0,995	0,993	NaiveBayes
Support Vector Machine	100	1,000	1,000	1,000	SMO
Trees	99.9812	1,000	1,000	1,000	Random Forest
	99.9906	1,000	1,000	1,000	J48
K Nearest Neighbor	99.79	0,999	0,999	0,999	IBK

4. Conclusion

In this research, a data set containing data related to the COVID-19 outbreak affecting the whole world and affecting everyone's life was used.

To this dataset, various classification algorithms were applied to "COVID-19 Mexico Patient Health Data Set" which was taken from the Kaggle site. The confusion matrix was used to calculate the performances of the classifiers.

Thanks to machine learning, the survival rate of patients was estimated in this data set consisting of 95839 cases. For this data set, it was witnessed that the support vector machine algorithm performs better than other classifiers.

When analyzing the features, it was shown that illnesses such as smoking, hypertension,obesity,diabetes,pneumonia are also being investigated for the patients. Factors affecting the healing status of patients may be explored in future studies by applying data mining to Covid-19 data. Additionally, the study may include several machine learning algorithms to make comparisons regarding classification success.

5. References

Agrawal., R. (2014). K-Nearest Neighborn for Uncertain Data. *International Journal of Computer Applications*, 13-16.

COVID-19 Mexico Patient Health Dataset. (2020, 05 19). Retrieved from Kaggle.com: https://www.kaggle.com/riteshahlawat/covid19-mexico-patient-health-dataset

Guan, W.-j. (2020). Clinical Characteristics of Coronavirus Disease 2019 in China. *New England Journal of Medicine*.

Ho, T. K. (1995). The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Hsu C. W., L. C. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 415-425.

Huang, C. (2020). Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *The Lancet*, 497-506.

L. Jiang, D. W. (2007). Survey of Improving Naive Bayes for Classification. *Lecture Notes in Computer Science*, 134-145.

McGoogan, Z. W. (2020). Characteristics of and important lessons from the coronavirus disease 2019 (covid-19) outbreak in china: summary of a report of 72 314 cases from the chinese center for disease control and prevention. *Jama*.

Quinlan, J. R. (1993). *Programs for Machine Learning.* Morgan Kaufmann Publishers.

Suchetana, B. R. (2017). Assessment of wastewater treatment facility compliance with decreasing ammonia discharge limits using a regression tree model. *Science of the Total Environment*, 249-257.