Journal of Engineering Technology and Applied Sciences, 2020 *e-ISSN:* 2548-0391
Received 20 October 2020, Accepted 29 December 2020 *Vol. 5, No. 3, 121-131*
Published online 30 December 2020, Research Article *doi: 10.30931/jetas.857665*

Citation: Çelik, A., "Using Apriori Data Mining Method in COVID-19 Diagnosis". Journal of Engineering Technology and Applied Sciences 5 (3) 2020: 121-131.

USING APRIORI DATA MINING METHOD IN COVID-19 DIAGNOSIS

Ahmet Çelik

Department of Computer Technology, Tavşanlı Vocational School, University of Kütahya Dumlupınar Türkiye
ahmet.celik@dpu.edu.tr

Abstract

Corona virus 2019 (COVID-19) disease has spread all over the world and many people have died due to this disease. PCR (Polymerase Chain Reaction) tests are mostly applied to detect people who have this disease. However, in some cases, it is necessary to wait twenty-four hours for the results of this test. In such cases, the treatment and isolation process of the patient may be delayed. Therefore, the rapid commencement of treatment and isolation process by analyzing the symptoms, are of great importance. Using data mining methods can be carried out quickly specify analysis. Association rule algorithms are also among data mining methods. The most common SETM, AIS and apriori association rule algorithms are encountered. The most widely used is the apriori association algorithm. Using this algorithm, the frequency and association rates of the data are found in the data set. In this study, it has been shown that association rules calculated by apriori algorithm can be used in the diagnosis of COVID-19. By using the COVID-19 Survilance data set, the association rates of the disease symptoms specified in the ICD (International Classification of Diseases) International Classification of Diseases codes were determined. According to the results obtained; it has been observed that the patients with these symptoms are 100% definitely infected with COVID-19 disease when the disease symptoms represented by the A01, A02 and A04 disease codes are together.

Keywords: Corona virus, COVID-19, information system, data mining, apriori algorithm

1. Introduction

Nowadays, with the development of technology, a large amount of data records have started to be formed by companies. Data create a large amount of sizes in many areas such as patient records of hospitals, customer records of communication companies, customer and product

records of shopping companies, product records of industrial manufacturing companies, records of security companies, personnel records. Large amounts of data should be analyzed and useful ones should be filtered. Data mining is used in many areas. Machine learning can also be done by using classification and clustering methods with data mining. With the help of many algorithms, useful data can be analyzed over the data set and a decision can be made about the new data. The data on data set can be divided into meaningful clusters or classes. Coronavirus is a large family of viruses that cause different diseases with a variety of symptoms. Most common are types of Coronavirus that have symptoms such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). However, Coronavirus Disease 2019 (COVID-19) is a new type of disease that has never been previously identified in humans [1]. The disease was first seen in China and soon spread all over the world and turned into a Pandemic disease.

Educational institutions also work to learn the success rates of students. It is necessary to use data mining techniques methods to analysis or predict students' learning performance. The use of data mining in the field of education is called Educational Data Mining. In addition, machine learning algorithms and statistical techniques is used in educational activities [2].

In recent years, a large amount of structured, unstructured and semi-structured data has been produced by institutions around the world. All of these data are called big data. The need to manage big data produced by various sources has also been encountered in the health sector [3]. Big data are recorded on cloud systems and analyzed [4]. In some cases, there may be unwanted data in the data obtained from real World sources. Therefore, noise needs to be cleaned. Data mining methods are used for this method too [5].

Wiguna and Riana [1], in their study; they used the C4.5 decision tree algorithm, a data mining method, using the COVID-19 Survilance dataset. COVID-19 Survilance dataset was published on 24 April 2020. On this data set, there are seven examples of symptoms shown by 14 people. In the study, it has been shown that decision trees can be used in the diagnosis of COVID-19 with the help of these symptoms.

Chen et al. [6], in the study; he analyzed large data sets in the medical field by using certain thresholds with the Apriori algorithm. They used medical data on cloud systems on HADOP. There is a huge amount of data in the medical field and it is increasing day by day.

Laksito and Kusrini [7], in their study; they have shown that using data mining Decision Trees methods, it makes a 90% significant improvement in financial invoice issuance compared to manual invoicing time
Yin et al. [8], in the study; he used data mining methods and prediction model in mining field. The ratio of technical and economic data of mining enterprises is multidimensional. By analyzing these data, the cost and sales prices of the product can be determined. For the prediction model, the price of mineral products was determined by using an artificial nerve.

Wu and Zhu [9], in their study; they proposed a dynamic data mining method for processing sensor data. However, an important sign of the Internet of Things era is that sensor data is replacing artificially compiled. The most important step is how to extract valuable information and patterns from the large amount of data generated by sensors. A data mining model has been created with sensors that can be used in the dynamic change process. In the last stage, the physical system was trained by collecting and reviewing historical changes in sensor data.

Hirano and Tsumoto [10], in their study; they have used data mining methods on medical data. In the field of medicine, big database information about both patients and diseases is created day by day.

Ya et al. [11], in the study; he conducted studies on rational drug use in pediatric diseases. Three drug samples were used and results were obtained on their interactions with the Apriori association rule.

Gutub and Ahmed [12], in their study, they explained a facial recognition system called Hajj & Umrah Face Recognition Database (HUFRD) that can be used in the Hajj and Umrah season in Saudi Arabia. There are frequent cases of disappearances during Hajj and Umrah seasons due to a very dense crowd of people. In this study; it is aimed to find missing people by analyzing the faces of people from cameras, using data mining and data matching methods, in the areas that are visited frequently.

Alassaf and Gutub [13], in their study; they used AES, SPECK and SIMON encryption methods to increase security on the data, during remote monitoring of personal healthcare data and performed performance comparison of these methods. Today, by using the Internet of Things (IOT), healthcare signals can be remotely monitored and analyzed. However, the security level should be kept under control and at a high level.

2. Materials and methods

In this study, apriori algorithm is used in data mining association rule algorithms. In this method, the frequency of which object an object carries with it is determined. The first known international classification was made by the International Statistical Institute in 1893. A list was published by the World Health Organization (WHO) as ICD-6 in 1948 and the ICD-10 list was published in 1990. The ICD-11 list was published on June 18, 2018 and it is planned that all member countries will report according to this list in 2022. Table 1 shows the international classification of symptoms.

Table 1. A00-A09 Infectious Diseases [14].

Disease Code	Disease Name
A00	Cholera
A01	Typhoid and paratyphoid fevers
A02	Other salmonella infections
A03	Shigellosis
A04	Other bacterial intestinal infections
A05	Other bacterial foodborne intoxications
A06	Amoebiasis
A07	Other protozoal intestinal diseases
A08	Viral and other specified intestinal infections
A09	Other gastroenteritis and colitis of infectious

2.1 Coronavirus disease 2019 (COVID-19)

This disease, which has affected the whole world, is thought to be caused by the transmission of the SARS-CoV-2 virus to humans from animals sold illegally to the Huanan Seafood Wholesale Market of China. This disease is transmitted from humans to humans through droplets. The droplets emitted by individuals carrying this virus as a result of coughing or sneezing are transmitted by touching the hands of other people and bringing their hands to their mouth, nose or eyes. Not all individuals carrying this virus show symptoms. Therefore, it is difficult to identify sick people. In individuals with symptoms, the symptoms of the disease are on average 5-6 days. However, in some cases it appears to be prolonged for up to 14 days. This virus causes more fatal cases in individuals with chronic diseases and the elderly (over 65 years of age). It is still uncertain that animals are the main source of COVID-19 transmission. The World Health Organization (WHO) states that Coronavirus 2019 disease (COVID-19) is a pandemic, an unnatural disaster.

2.2 Apriori Algorithm

The Apriori algorithm was introduced by Aggrawal in 1993. This algorithm; It is widely used to create association rules between data. Two important parameters are taken into account when creating the rule. These parameters; confidence (precision value): strong connection and support (support value) between items in the data set: indicates combination percentage of data set items [7].

Formula 1 shows, the calculation of the support value in the data set.

$$Support(A > B) = \frac{Frequency(A,B)}{N} \quad (1)$$

Formula 2 shows, the calculation of the confidence value in the data set.

$$Confidence(A > B) = \frac{Frequency(A,B)}{Frequency(A)} \quad (2)$$

A and B are the clusters and the N value is the total number of data in the data set. The association rule can make predictions for the future using data mining algorithms the variety, placement and production quantities may change, considering the product demands of the customers. This methods increase the earnings of sales companies [16]. Repeating data structures between different data groups are revealed with the Apriori algorithm. It is an original algorithm that uses a forward, iterative approach. In this algorithm, the entire data set must be mapped [17].

The Apriori algorithm is also implemented on values in Matrix form. However, it reveals the situation of scanning the data repeatedly and loading the input/output processes repeatedly [18]. Apriori algorithm has a different functioning than AIS and SETM algorithms. In this algorithm, candidate object clusters are not produced during scanning, but operations are performed on candidate clusters that were detected to be large in the previous scan [19].

Searching frequently used items in a data set with the Apriori algorithm can take time. Therefore, efforts are being made to increase the speed of the Apriori algorithm [20].

3. Findings and discussion

Machine learning is the ability of the machine to make decisions in the new situation, using the data in the data set which is given as an experience. In machine learning, data set data is a very important source of experience [21]. The data of this data set is shown in Table 2. In the data set the disease symptoms of people with corona virus, people in monitoring and person without symptoms are shown. PUS (Patient Under Supervision) refers to patients under surveillance, PIM (Person in Monitoring) to people under observation (suspected person) and PWS (Person without Symptoms) to people who show no symptoms [22]. In this study, the data set was analyzed assuming Support 60% and Confidence 75% using the Apriori algorithm. Thus, it is aimed to achieve high reliability.

Table 2. COVID-19 Surveillance data set data [22].

A01,A02,A03,A04,A05,A06,A07,Categories
+,+,+,+,+,-,-,PUS
+,+,-,+,+,-,-,PUS
+,+,+,+,-,+,-,PUS
+,+,-,+,-,+,-,PUS
+,-,-,-,-,-,+,PUS
+,+,+,-,-,-,+,PUS
+,+,-,-,-,-,+,PUS
+,+,+,+,-,-,-,PUS
+,-,-,+,+,-,-,PIM
-,+,-,+,+,-,-,PIM
+,-,-,+,-,+,-,PIM
-,+,-,+,-,+,-,PIM
-,+,-,-,-,-,+,PIM
-,-,-,-,-,-,+,PWS

The diseases specified with ICD codes in the data set are shown in Table 3 in Boolean (True/False) format. The ‘+’ symbol indicates that the ICD coded disease is present (true: 1) and the ‘-’ symbol indicates that the ICD coded disease is absent (false: 0).

Tablo 3. Display of diseases in Surveillance 2020 dataset in Boolean (True / False) format.

A01	A02	A03	A04	A05	A06	A07	Categories
1	1	1	1	1	0	0	PUS
1	1	0	1	1	0	0	PUS
1	1	1	1	0	1	0	PUS
1	1	0	1	0	1	0	PUS
1	0	0	0	0	0	1	PUS
1	1	1	0	0	0	1	PUS
1	1	0	0	0	0	1	PUS
1	1	1	1	0	0	0	PUS
1	0	0	1	1	0	0	PIM
0	1	0	1	1	0	0	PIM
1	0	0	1	0	1	0	PIM
0	1	0	1	0	1	0	PIM
0	1	0	0	0	0	1	PIM
0	0	0	0	0	0	1	PWS

The single frequency of Covid-19 symptoms for PUS, PIM and PWS decisions is shown in Table 4. The number of patient with A01(8) symptom is the highest. The number of patients with A05 symptom and the A06 symptom is the lowest. None of the people showed A03 symptom under observation. The person who only shows A07 symptom is not ill.

Table 4. Frequency of A01-A07 symptoms alone.

Frequency	PUS	PIM	PWS
A01	8	2	0
A02	7	3	0
A03	4	0	0
A04	5	4	0
A05	2	2	0
A06	2	2	0
A07	3	2	1

The frequency of the disease symptoms in pairs is shown in Table 5. In patients with COVID-19 virus, it was observed that A01(7) and A02(7) disease symptoms were mostly together.

Table 5. Frequency of coexistence of two symptoms.

ICD Codes		Frequency Values
A01	A02	7
A01	A03	3
A01	A04	5
A01	A05	2
A01	A06	2
A01	A07	3
A02	A03	3
A02	A04	5
A02	A05	2
A02	A06	2
A02	A07	3
A03	A04	3
A03	A05	1
A03	A06	1
A03	A07	1
A04	A05	2
A04	A06	2

In this study; it is aimed to determine the rule of concomitance of three disease states with the Apriori algorithm, Table 6 shows the coexistence of three symptoms. These symptoms are data from patients under surveillance.

Table 6. Symptoms of supervised patients.

PUS-ICD Codes						
A01	A02	A03	A04	A05		
A01	A02		A04	A05		
A01	A02	A03	A04		A06	
A01	A02		A04		A06	
A01						A07
A01	A02	A03				A07
A01	A02					A07
A01	A02	A03	A04			

Five situations with three symptoms of illness and the levels of Support, Confidence of these situations are shown in Table 7. According to this; In all five cases, Support values were calculated at 62.5%. Case 3: If A04 is present, then A01 and A02 exist. Case 5: If A01 and A04 are present, then A02 is present. The confidence level for Case 3 and Case 5 association rules is calculated as the highest (1: 100%).

Table 7. Coexistence of three symptoms.

Disease Coexistence of Patients Under Supervision		Support (%)	Confidence (%)
Case 1	If A01(8) is present, A02 and A04(5) are exist together.	62,5	62,5
Case 2	If A02(7) is present, A01 and A04(5) are exist together.	62,5	71,4
Case 3	If A04(5) is present, A01 and A02(5) are exist together.	62,5	100
Case 4	If A01 and A02(7) are exist together A04(5) is present.	62,5	71,4
Case 5	If A01 and A04(5) are exist together A02(5) is present.	62,5	100

In this study; as success targets were determined initially 60% Support and 75% Confidence levels. The support level for all cases is calculated as 62.5%. This value is higher than the initial target. Confidence level 71% was calculated for Case 2, Case 3, Case 4 and Case 5. This is lower than the initial target. However, Confidence level for Case 3 and Case 5 is 100%. It appears that these two conditions can be used in the diagnosis of COVID-19, as the level of accuracy is very high. The Support and Trust levels of all situations are shown on figure 1.

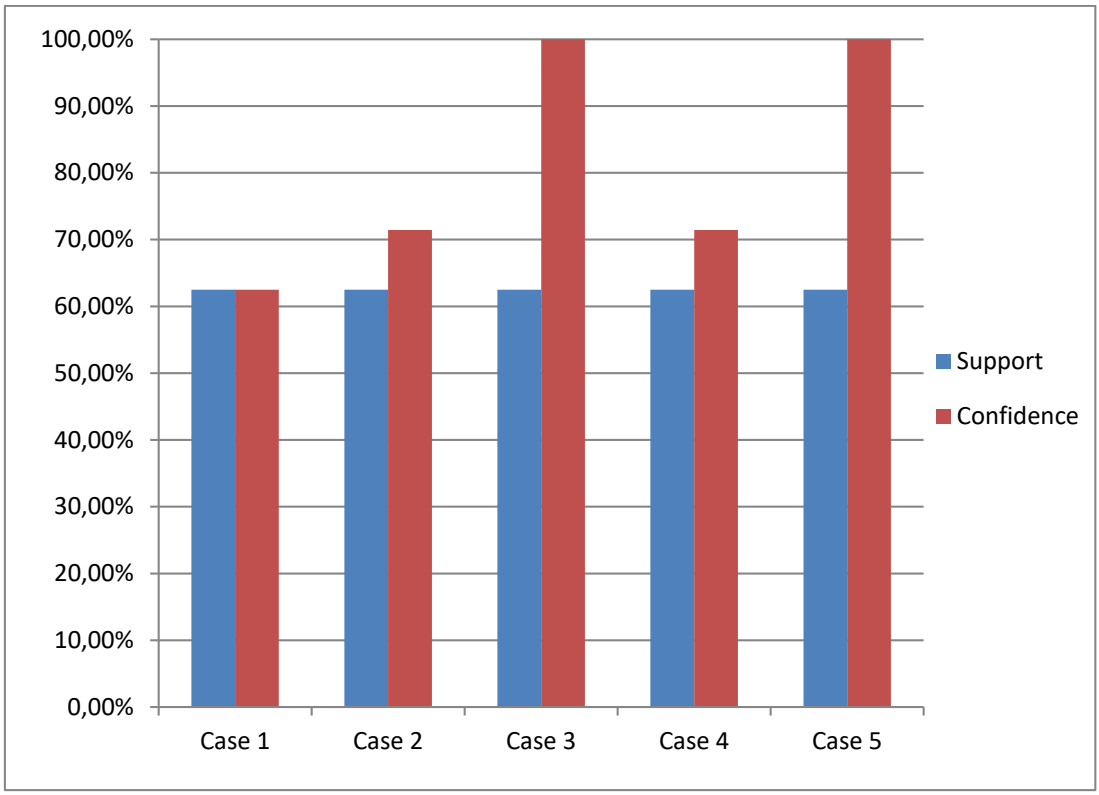

Figure 1. Support and Trust levels calculated by Apriori Algorithm

In addition, the data of persons under observation (PIM) is included in the COVID-19 Survilance data set. Data on persons under observation are given in Table 8. These people are not accepted as a patient under observation. Their situation is monitored by being kept under observation. When the data were analyzed, none of the subjects under observation showed A03 symptoms. In addition, A01 and A02 disease symptoms were never seen together.

Table 8. Symptoms of people under observation.

PIM-ICD Codes						
A01			A04	A05		
	A02		A04	A05		
A01			A04		A06	
	A02		A04		A06	
	A02					A07

It is very unlikely that three conditions will occur in people under observation. The most common A04 symptom is seen.

4. Results

In this study, data mining method was applied on the symptoms of COVID-19 disease which turned into a pandemic epidemic. Apriori algorithm is used as data mining method. COVID-19 Surveillance data set was used. This dataset was published on 24 April 2020. There are fourteen people's symptoms in the data set. This data set includes some disease ICD codes for diagnosis COVID-19 that is called Patient Under Supervision, Person in Monitoring and Person Without Symptoms.

In the study; association status of the individuals was examined with the Apriori algorithm on the disease symptoms and the most common symptoms were shown in the data set. In the diagnosis of COVID-19 disease, it has been observed that there are symptoms of the disease specified in codes A01 to A07. It has been observed that “if A04 is present, A01 and A02 are both together exist” and “if A04 and A01 are both together exist, A02 is present” in 100% of the patients. That is "if people have Bacterial Intestinal Infection, Typhoid and Paratyphoid and Salmonella infection are together", then the person is COVID-19 positive. In addition, “if people have Bacterial Intestinal infection and Typhoid - Paratyphoid are both together, Salmonella infection is present", then the person is COVID-19 positive. In addition, the persons who have only with A07 symptoms that is not patients. It is necessary to wait twenty-four hours for the PCR test results used to detect COVID-19 disease. In this study, it is shown that the disease can be diagnosed much faster by evaluating the coexistence of the symptoms. Thus the treatment and isolation process of the patients will be performed faster.

References

[1] Wiguna, W., Riana, D., “Diagnosis of Coronavirus disease 2019(Covid-19) surveillance using C4.5 algorithm”, Journal PILAR Nusa Mandiri, 16 (2020) : 71-80.

[2] Jalota, C., Agrawal, R., “Analysis of educational data mining using classification”, 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con); 14th -16th Feb 2019; India. (2019) : 243-247.
[3] Kumar S., Singh, M., “Big data analytics for healthcare industry: impact, applications and tools”, Big Data Mining and Analytic, 2 (2019) : 48-57.
[4] Kotturu, P.K., Kumar, A., “Data mining visualization with the impact of nature inspired algorithms in big data”, Proceedings of the Fourth International Conference on Trends in Electronics and Informatics (ICOEI 2020); 16-18, April 2020. India. Tirunelveli, (2020) : 664-668.
[5] Cui X., Yang, S., Wang, D., “An algorithm of apriori based on medical big data and cloud computeing”, 3th Conferance of Computational Interdisciplinary Science; 7-10 November 2016; Campos, Brazil, (2016) : 361-365.
[6] Chen, Y.C., Chang, Y.T., Kan, Y.S., Chen, R.S., Wu, S.F., “Using data mining technique to improve billing system performance in semiconductor industry”, International Conference on Information and Computer Technologies; 23-25 March 2018, Dekalb, United States, (2018) : 1-4.
[7] Laksito, A.D., Kusrini, K., “Apriori algorithm optimization using temporary table”, International Conference on Information and Communications Technology (ICOIACT); 24-25 July 2019. Yogyakarta, Indonesia, (2019) : 560-565.
[8] Yin, Y., Long, L., Deng, X., “Dynamic data mining of sensor data”, IEEE Access, (2020) : 8:41637-61648.
[9] Wu, X., Zhu, X., “Mining with noise knowledge: Error-aware data mining”, IEEE Transactions on Systems Man and Cybernetics, 38 (2008) : 917-932.
[10] Hirano, S., Tsumoto, S., “Frequent temporal pattern mining for medical data based on ranged Relations”, IEEE International Conference on Data Mining; 18-21 November 2017. New Orleans, United States (2017) : 612-616.
[11] Ya, L., Lei, Y., Li, W., “Chun M, Guiming Y. “Application research of apriori algorithm based on matrix multiplication in children’s drug interaction”, 12’th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA) 2020; 28-29 Feb. 2020. Phuket Thailand, (2020) : 507-512.
[12] Gutub, A., Ahmed, S., "Trialing a smart face-recognition computer system to recognize lost people visiting the two holy mosques", Arab Journal of Forensic Sciences & Forensic Medicine (AJFSFM), 8 (2019) : 1120-1133.
[13] Alassaf, N., Gutub, A., "Simulating light-weight-cryptography implementation for IoT healthcare data security applications", International Journal of E-Health and Medical Communications (IJEHMC), 4 (2019) : 1-15.
[14] World Health Organization (WHO), “International statistical classification of diseases and related health problems”, 10’th Revision (ICD-10 manual), Centers for Disease Control and Prevention, 1 (2005) : 1-1268.
[15] T.C. Sağlık Bakanlığı Halk Sağlığı Genel Müdürlüğü, “COVID-19 (SARS-Cov-2 Enfeksiyonu rehberi) Bilim Kurulu Çalışması”, 12 Nisan 2020, Ankara, (2020) : 1-98.
[16] Zerman, M., “Birliktelik kuralı algoritmaları ile büyük veriler üzerinde analitik analizler: havaalanı örneği”. Yüksek Lisans Tezi, Haliç Üniversitesi, İstanbul, Tükiye, (2018).

[17] Singh, J., Ram, H., Sodhi, J.S., “Improving efficient apriori algorithm using enhanced transaction reduction”, International Journal of Scientific and Research Publications, 3 (2013) : 1-4.
[18] Wang, F., LI, Y., “An improved Apriori algorithm based on matrix”, 12’th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA) 2020. 28-29 Feb. 2020. Phuket, Thailand, (2020) : 488-491.
[19] Silahtaroğlu, G., “Veri madenciliği (Kavram ve algoritmaları)”, 3. Basım, İstanbul, Türkiye: Papatya Yayıncılık Eğitim, (2016).
[20] Karimtabar, N., Shayegan Fard M.J., “An extension of the apriori algorithm for finding frequent items”, 2020 6’th International Conference on Web Research (ICWR) 2020. 22-23 Apr. 2020. Tehran, Iran, (2020) : 330-334.
[21] Balaban, M.E., Kartal, E., “Veri madenciliği ve makine öğrenmesi temel algoritmaları ve R Dili ile Uygulamalar”, 2. Basım, İstanbul, Türkiye: Çağlayan Kitap & Yayıncılık&Eğitim, (2018).
[22] Dua, D., Graff, C., “UCI machine learning repository”, University of California, School of Information and Computer Science. Irvine, USA, (2019).