A Novel Approach of Frequent Itemsets Mining for Coronavirus Disease (COVID-19)

Mai Shawkat, Mahmoud Badawy, and Ali I. Eldesouky

Abstract **— The global pandemic of new coronaviruses (COVID-19) has infected many people around the world and became a worldwide concern since this disease caused illness and deaths. The vaccine and drugs are not scientifically established, but patients are recovering with antibiotic drugs, antiviral medicine, chloroquine, and vitamin C. Now it is obvious to the world that a quicker and faster solution is needed for monitoring and combating the further spread of COVID-19 worldwide, using non-clinical techniques, for example, data mining tools, enhanced intelligence, and other artificial intelligence technologies. In this paper, association rule mining is developing for the frequent itemsets discovery in COVID-19 datasets, and the extraction of effective association relations between them. This is done by demonstrates the analysis of the Coronavirus dataset by using the Apriori_Association_Rules algorithm. It involves a scheme for classification and prediction by recognizing the associated rules relating to Coronavirus. The major contribution of this study employment determines the effectiveness of the Apriori_Association_Rules algorithm towards a classification of medical reports. The experimental results provide evidence of the Apriori_Association_Rules algorithm regarding the execution time, memory consumption, and several associated rules that reflect its potential applications to different contexts. Therefore, the Apriori_Association_Rules algorithm will be very useful in healthcare fields to demonstrate the latest developments in medical studies fighting COVID-19.**

Index Terms **— COVID-19; Coronavirus; Classification; data mining; pattern matching.**

I. Introduction

The 2019 Coronavirus disease (COVID-19) has belonged to a large Coronavirus family (CoV). The onset of serious disease may lead to death due to it invades our respiratory cells and triggers an immune response that targets those infected cells, destroys the lung tissue, and eventually clogs our airways, cutting off our supply of oxygen [1]. Early and automatic diagnosis can be useful for countries to rapidly refer the patient to quarantine, to a quick insight of severe cases in specialist hospitals, and to monitor disease spreading. As the diagnosis is an extremely quick process, the financial problems caused by the cost of additional studies affect both states and patients, particularly in regions with private health systems, or because of financial constraints restricted access to medical systems. Machine learning has developed disciplines to apply artificial intelligence to mine, analyze, and identify patterns of data. The preservation of developments in these fields for the benefit of medical decision-making and computer-aided application has become an important need as new data exists [2].

Data mining is generally concerned with classification, clustering, regression, and correlation. Data mining generates information about formerly accurate independent itemsets and their connection in large datasets. The represented items are called frequent itemsets. Extracting frequent itemsets from the transaction databases is intended to the Frequent Itemset Mining. It is important for exploring association rules widely used in different real-life applications such as web data analysis, consumer behavior analysis, cross-marketing, catalog design, and medical records.

As well as Association Rule Mining (ARM) involves in biomedical sciences and investigates for providing prediction for disease detection and accurate classification [3]. At a basic level, the ARM includes using data mining methods to analyze patterns in data, or correlation, within a dataset. Association rules consist of if-then statements that represent the possibility of relationships between itemsets in big databases with different types. This defines every if-then association that is referred to as association rules [4]. ARMis generated by looking for recurrent data patterns while using the support and confidence parameters to find relationships between apparently unrelated datasets or another data repository. Support seems to be an indicator of the frequency of relationships that appears in the database, while confidence shows how many times these relationships proved to be true.

Many algorithms for mining the association rules are discussed elaborately and the performance can be seen in the literature survey of ARM [5]. The traditional algorithm Apriori [6] discovers frequent itemsets from big transactional databases. It is a later stage it extends them to larger item sets that occur frequently. The frequent itemsets found by Apriori are used to explore association rules. This can be used in the application domain market basket analysis. In the mining process, the Apriori algorithm uses dataset scanning and pattern matching to calculate the frequency of the candidate itemsets. Many optimization algorithms [7], [8] were presented in the literature, based on the original algorithm, to increase the effectiveness of Apriori.

Submitted on February 20, 2021.
Published on March 15, 2021.

Mai Shawkat, Department of Computer Engineering and Systems, Faculty of Engineering, Mansoura University, Mansoura, Egypt.
(corresponding e-mail: shawkatmai@gmail.com)
Mahmoud Badawi, Department of Computer Engineering and Systems, Faculty of Engineering, Mansoura University, Mansoura, Egypt.
(e-mail: engbadawy@mans.edu.eg)
Ali I. Eldesouky, Department of Computer Engineering and Systems, Faculty of Engineering, Mansoura University, Mansoura, Egypt.

DOI: http://dx.doi.org/10.24018/ejece.2021.5.2.304

The FP-growth is a famous algorithm of infrequent patterns mining [9], [10]. It relies on a prefix-tree configuration used to store a database in that compact form (FP-tree). In this strategy, the tree is divided into sub-trees called conditional FP-trees using a dataset called conditional pattern base. This recursive mining task affects the time consumed during the mining process. To eliminate this deficiency, this paper based on the existing techniques sets out an Apriori_Association_rules algorithm.

However, in medical diagnosis, the ARM can be useful for physicians to treat patients. Diagnosis is not a simple task and has some errors that can lead to inaccurate results. The probability of the disease arising in different factors and symptoms can be defined using relational association rule mining. Further, this framework can be developed using learning methods, introducing new signs, and identifying relationships between these new signals and their disease.

Monitoring a large number of reported cases for successful isolation and treatment is a priority to manage the transmission of Coronavirus disease. Pathogenic laboratory testing is the significant standard research, but it is time-consuming despite its precise results. Quick and accurate methods of diagnosis are urgently required for combating the disease.

The Apriori algorithm is proposed not only for reducing runtime and memory usage but also for decreasing the overall cost. An analysis is used to see which feature arrays/attributes are often useful in distinguishing different classes. In association analysis methods, the discovery of behavioral patterns was also widely used, mainly with support and confidence to determine if there is any relationship among data. The concept is to find frequent items first according to the level of support by searching the database, and then generate association rules referring to the level of confidence [11]. The possible associations can be found in a big database, based on the ARM between items [12].

The rest of the paper is organized as follows. In the related work section, some of the existing data mining algorithms in healthcare applications were reviewed. The methodology section depicts the Apriori_Association_Rules algorithm. The experiments section exhibits the comparative experiments using real COVID-19 datasets, includes: (1) dataset description (2) runtime consumption (3) memory consumption (4) Number of Association Rules. Finally, the Conclusion section ends the paper with results and some future work.

II. Related Work

Data mining is the study of recognizing hidden patterns and relationships in big databases that were entirely unknown [13]. Data mining techniques were commonly used in many applications in healthcare such as simulation of clinical outcomes and patient predictions, evaluation of therapy efficiency, the ranking of hospitals, and monitoring of infections [14].

The current Coronavirus study is dedicated to providing stronger surveillance to limit the effect of Coronavirus outbreaks. Scientific research involves a wide range of studies aimed at learning how the infection is caused and how it destroys causing disease. Throughout a general direction, an approach to the association rules in the Coronavirus was proposed to diagnose Coronavirus symptoms. The major studies referring to pattern recognition in time series and biomedical sequences are listed below.

Venkatalakshmi and Shivsankar [15] examined the efficiency of the Naive Bayes decision tree algorithm for heart disease diagnoses. Naive Bayes, multi-layer perception, and basic logistics have developed the best predictive models. The experimental results used 294 reports with 13 attributes measuring the two algorithms' efficiency in the database. For breast cancer diagnosis and prognosis, FP-growth algorithm, and association rule mining were applied [16]. The classification models have been constructed to use 699 records and 9 attributes dataset, as well as the highest accuracy, which is being achieved with decision trees induction algorithms.

Bellaachia et al [17] used Naive Bayes for the decision tree to predict breast cancer survival patients. Several predictive models have been developed recently for breast cancer survivals [18]. Three different techniques have been employed for data mining: The Support Vector Machine (SVM), Chi-squared Automatic Interaction Detection (CHAID), and Bayes Networks, and it found that SVM was the best survival prediction model [19], [20].

Frequent itemset mining procedures are performed on various computing nodes on a distributed system and obtain final results by collecting and analyzing contextual outputs. Another popular distributed computing procedure Map-Reduce [21] implemented at Hadoop that provides a reliable, accurate, and defect tolerance tool for large volumes of data. For example, PFP [22] provides three Map-Reduce stages in which FP-growth tasks can be separated and intermediate outputs combined. For frequent itemset extraction, the data and the tasks are parallelized. PFP achieves a higher efficiency as regards time and scalability. The Map-Reduce method is used to perform mining tasks through the creation of an early stage boundary for extracting infrequent itemsets in the FNBP node sets algorithm [23]. Experiments have measured its workload and scalability balance performance. However, because of frequent input/output operations during the entire computing process, Map-Reduce would not be useful for implementing workflows that are extremely common in ARM algorithms.

YAFIM [24] has been proposed to improve the Spark memory-based workflow engine, due to increased use of the Apache Spark Platform [25]. DFIMA was proposed in [26] by Zhang et al. to reduce multiple computations in the pattern growth phase, using matrix pruning approaches. On another hand, PAPT-growth [27] enables large amounts of data to be processed as a Spark workflow.

The AC method [28] also exists, but it suffers from various regulations. CBA is one of the AC developed algorithms, as well as it uses the function of generating Apriori candidates to discover new rules of the association from datasets. An improved Apriori algorithm (EAA) for a sequential minimal optimization (SMO) is proposed based on the knowledge of context ontology (EAA-SMO) [29] showed an increase in the accuracy of the whole process. Recently, the Intelligent Apriori (lAP) algorithm [30] uses an extension of the Apriori algorithm for the mining and processing of the data obtained using patterns and relationships, using the frequent itemsets.

However, the efficiency of the algorithm must be enhanced. The primary contribution of this paper is to develop the Apriori algorithm by improving the efficiency of runtime and memory consumption of the algorithm and association rules mining process that is useful for managing and classification of COVID-19 data sets.

III. The Proposed Approach

By mining COVID-19 data sets, the organizations involved hope to construct advanced mapping tools capable not only of monitoring the current spread of the virus but also of predicting how it will develop in the future. The proposed Apriori_Association_Rules algorithm aims to reduce execution time and memory consumption with just one database scan and improves the effectiveness of running time and storage of data structures.

However, the Apriori_Association_Rules tree consists of the node layout prefix tree with four fields and the frequently reported candidates header table with two: one field holds to the item, while the other field holds the frequency total counts for the item. The data between itemsets is also stored. The Apriori_Association_Rules uses frequent candidates to produce association rules. This proposed algorithm traverses the tree way quicker. The pseudo-code of the proposed algorithm is presented as follows.

Pseudo-code:

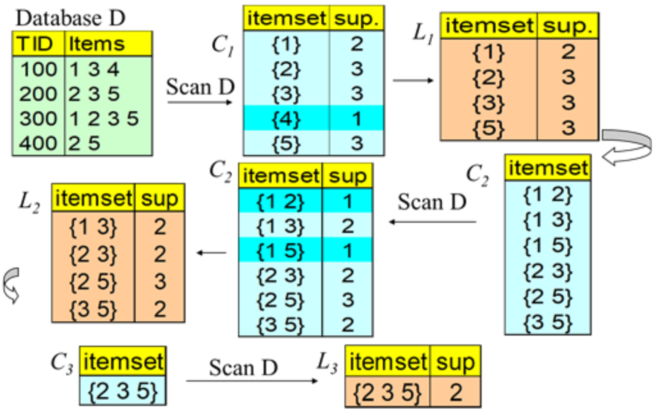
```
$C_m$: Candidate itemsets of  m
$L_m$ : frequent itemsets ofm
$L_1$ = {candidate itemset};
for (m = 1; $L_m$ !=∅; m++) do begin
    $C_{m+1}$ = candidates created from $L_m$;
for each transaction k in dataset do
      increment the count of all candidates in $C_{m+1}$
Which are included in k
    $L_{m+1}$  = candidates in $C_{m+1}$ with min_sup
 end
return ∪$_m$ $L_m$;
```

Overall, the output candidates can be generated by:

```
•  Assume the items in $L_{m-1}$ are registered  in order
•  Step 1: join $L_{m-1}$
        Insert into $C_m$
        Select p.item$_1$, p.item$_2$… p.item$_{m-1}$, q.item$_{m-1}$
        From $L_{k-1}$ p, $L_{k-1}$ q
        Where  p.item$_1$=q.item$_1$…  p.item$_{m-2}$=q.item$_{m-2}$,
p.item$_{m-1}$<q.item$_{m-1}$
•  Step 2: pruning
        For all item set c in $C_m$ do
        For all (m-1)-subset s of c do
            If (s is not in $L_{m-1}$) then delete c from $C_m$
```

Each transaction contains several candidates which makes the total number of candidates huge. Candidate items are classified in a hashed array (HAT) invented by Edward Sitarsk [31]. The leaf node of the hashed tree includes itemsets and counts. The internal node includes the hash table. The subset function discovers all the candidates involved in a transaction. For each item to be collected and scanned in the current Hash-based itemset counting to gather the conditional pattern base for that item, the Apriori algorithm uses frequent (m-1) itemsets to create candidates repeatedly, without the optimum use of certain pruning strategies. The pattern growth was guided primarily by database scanning and pattern matching to calculate the candidate counts.

Similarly, through the operations mentioned above, the construction process for the Apriori_Association_rules algorithm is explained in the example in Fig. 1. Items with lesser support value are pruned.



Database D

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

itemset	sup.
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

itemset	sup.
{1}	2
{2}	3
{3}	3
{5}	3

itemset
{1 2}
{1 3}
{1 5}
{2 3}
{2 5}
{3 5}

itemset	sup
{1 2}	1
{1 3}	2
{1 5}	1
{2 3}	2
{2 5}	3
{3 5}	2

itemset	sup
{1 3}	2
{2 3}	2
{2 5}	3
{3 5}	2

itemset
{2 3 5}

itemset	sup
{2 3 5}	2

Fig. 1. The Apriori Algorithm Example.

IV. Experimental Results

The Apriori_Association_Rules algorithm is used to reprocess information provided by datasets for efficient data analysis and finding patterns and rules. The first experiment conducts the runtime of the Apriori algorithm. The second experiment tests the memory. The third experiment mine association rules for Covid-19 datasets to verify our proposed Apriori_Association_Rules algorithm. Different support and confidence levels for each dataset are used. The experimental setup was performed on a Dell -Inspiron 7th Generation Intel® Core™ i7-7200U with 8 GB DDR4 2400MHz memory, 3M Cache, up to 3.10 GHz, AMD Radeon™ R5 M430 4GBand 1TB hard drive running Windows 10.

A. Datasets

The first dataset called COVID-19-inside-Hubei [32] reports patients inside the Chinese area of Hubei. The second dataset is called COVID-19-outside-Hubei [32] including the patients outside of the Hubei Province of China. The third dataset, named COVID-19-Merging [33]. This dataset is generated by a combination of data sets from different countries to generate a large dataset. Lastly, the fourth dataset is named DS4C (Data Science for COVID-19) dataset [34]. It is a structured dataset contained the reports of KCDC and regional governments.

B. Runtime Consumption

This experiment uses numerous datasets of different sizes as the input with multiple confidence levels to verify the time performance of the proposed Apriori_Association_Rules algorithm. COVID-19-inside-Hubei, COVID-19-outside-Hubei [32], and COVID-19-Merging [33] are used for four

DOI: http://dx.doi.org/10.24018/ejece.2021.5.2.304

separate sets of results. The consuming time required to process a particular dataset is shown in Fig. 2.

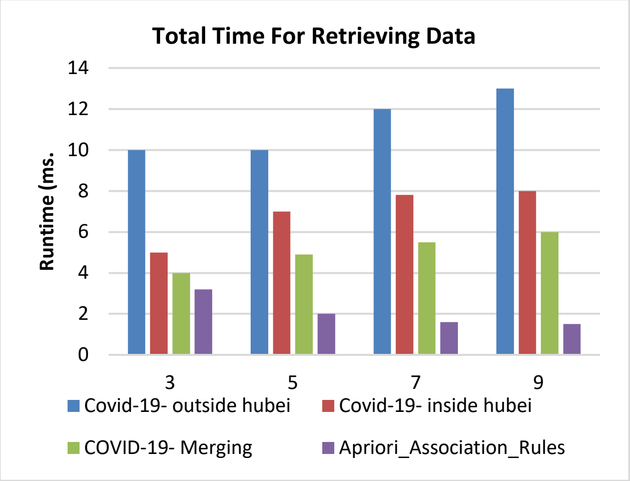


Fig. 2. The runtime required for processing data on the COVID-19 data sets.

From Fig. 3 y-axis display runtime in millisecond and x-axis display the different minimum confidence values. It includes the results of the execution time of the proposed Apriori algorithm and Covid-QF framework [35] on the COVID-19-inside-Hubei, COVID-19-outside-Hubei [32], and COVID-19-Merging [33] datasets with different minimum confidence levels. The graph is showing clearly that the Apriori_Association_Rules algorithm does well in comparison with COVID-QF. It is because the proposed algorithm pruning procedure discards the generation of conditional sub-trees. Thus, the runtime to mine frequent itemset is faster.

We conduct the experiment on DS4C (Data Science for COVID-19) dataset [34] from the Kaggle repository.

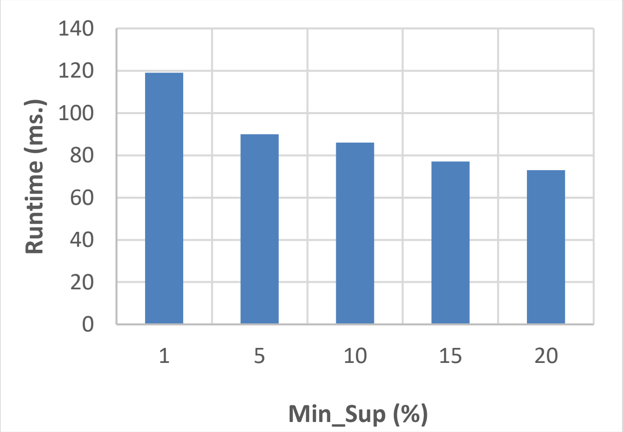
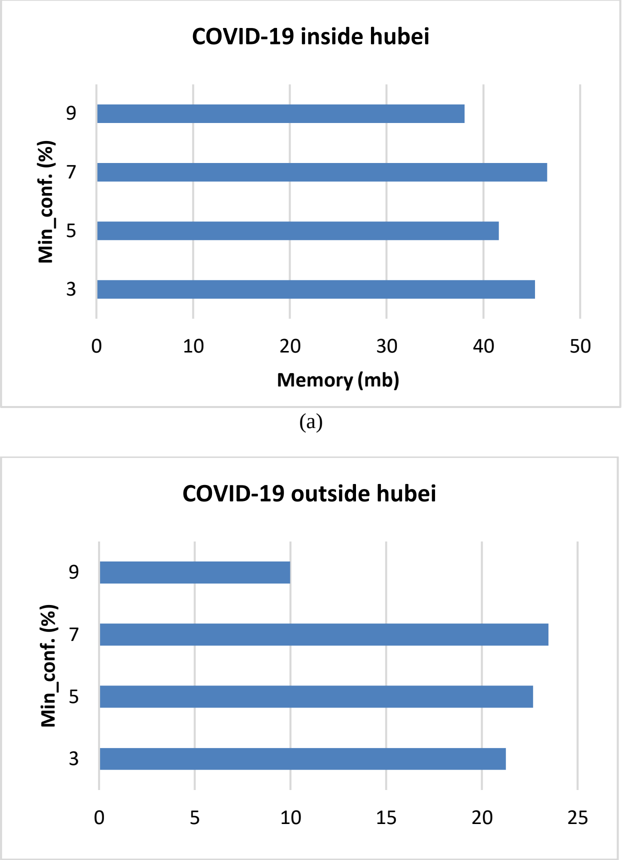


Fig. 3. Running time performance.

The time needed for the algorithm to construct the classifier on the covid-19 dataset is portrayed in Fig. 3. The runtime of both algorithms varies with different minimum support levels. Due to the Apriori pruning procedure discards the generation of conditional sub-trees, the time consumption is very low which makes the mining speed so fast.

C. Memory Consumption

The memory in MB is shown from the graph longitudinal axis, and the different confidence values are shown from the latitude axis. The charts depicted in Fig. 4 show the memory utilization of the proposed algorithm Apriori_Association_Rules. As the proposed algorithm takes the bottom-up technique, conditional pattern base generation and conditional subtree are not required. Thus, the memory for storing conditional patterns is avoided.



(a)

(b)

Fig. 4. The memory consumption on the COVID-19 data sets.

Clearly, from the results, the Apriori_Association_Rules algorithm is a useful technique for mining the maximum frequent itemsets, specifically for dense datasets. The results obtained are summarized in Table I.

We conduct the experiment on DS4C (Data Science for COVID-19) dataset [34] from the Kaggle repository to determine the memory consumption needed with the Apriori_Association_Rules algorithm.

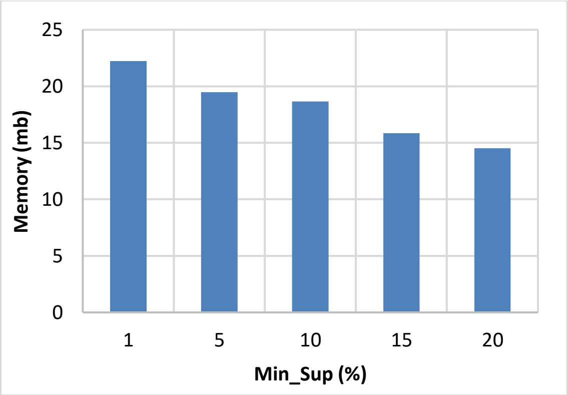


Fig. 5. Memory consumption.

The memory consumption of the proposed Apriori_Association_Rules algorithm is 23MB for the minimum support threshold level of 1% for the DS4C dataset as shown in Fig. 5. The creation of a conditional pattern base and conditional subtree is not needed as the proposed algorithm takes the bottom-up approaches. This avoids the memory needed to store conditional patterns.

D. Association Rule Mining

Association rules reveal a relationship in datasets between items. Association rules derived from medical reports can help to find a connection between different diseases, diseases and symptoms, diseases, and medicines. The core idea is to discover the most interesting and frequent correlations, called associations, in a considerable amount of data. This experiment mines association rules for Covid-19 datasets [32] [33] [34] to verify our proposed Apriori_Association_Rules algorithm. It can be accomplished by analyzing rules that can show an aspect or subsequence occurring. The frequencies in the dataset are determined in each sequence to get an association rule. The rule can be represented as A→C, where A, C are the relevant elements.

TABLE I: MEMORY AND RUNTIME COMPARISON USING APRIORI

Dataset min_conf (shown as δ %)	Algorithm	Memory (MB)			Runtime (ms)			Channels
		δ1	δ2	δ3	δ_1	δ_2	δ_3	Main channel
COVID-19-inside-hubei δ1=3, δ2=5, δ3=7, δ4=9	Apriori_Association_Rules	118	100	120	160	86	116	Assistant channel
COVID-19-outside-hubei δ1=3, δ2=5, δ3=7, δ4=9	Apriori_Association_Rules	119	120	100	138	149	158	
COVID-19-Merging δ1=3, δ2=5, δ3=7, δ4=9	Apriori_Association_Rules	102	115	116	1100	1180	1050	

The support threshold and the confidence threshold are affected by the consistency and quantity of the established rules. Table II has shown the results of the association rule mining of the COVID-19-inside-Hubei, and COVID-19-outside-Hubei datasets.

TABLE II: NO. OF ASSOCIATION RULES

Min_conf (%)	COVID-19-inside-Hubei Association rules no.	COVID-19-outside-Hubei Association rules no.
3	10275	9129
5	10201	9085
7	10186	8812
9	10186	8682

Table III identifies the elements used to present basic statistical indicators and to search for rules of the association. The number of association rules of the DS4C (Data Science for COVID-19) dataset generated by the Apriori algorithm is presented in Table III.

TABLE III: COLLECTION OF ATTRIBUTES SEEN WITHIN ASSOCIATION RULES

Item	Description
Case_id	The ID of the infection case
Province	City/ Metropolitan city/ Province (-do)
Age	The age of the patient
Sex	The sex of the patient
City	City/country/district
Group	True: group infection False: not group
Infection-case	The infection case (the name of a group or other cases)
#confirmed	The count of confirmed cases
Latitude	The latitude of the group
Longitude	The longitude of the group

Table IV presents the number of association rules generated by using the proposed method under different minimum support levels.

TABLE IV: NO. OF ASSOCIATION RULES GENERATED BY APRIORI ALGORITHM

Min_Sup (%)	No. Of Association Rules
1	287
5	22
10	3
15	2
20	2

The result of the algorithm Apriori is an association rule, such as one with this form:"attribute_0=239 ==> "id","age","sex","city","province","country","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","admin3","admin2","admin1","country_new","admin_id"="NA"",1.0,1.0

The Apriori_Association_Rules algorithm task is to search for association rules from a given set. This algorithm was also used through the Rules Package in our analysis of association rules. Apriori discovered the association rules in two stages: classifying the sets of items and generating the association rules. Figure 6 shows a specific association rule from an illustrative set of data that can be interpreted. The result of the algorithm according to the selected parameters is shown in Fig. 7.

```
Pattern,#SUP:,#CONF:
"attribute_0=1 ==>
"id","age","sex","city","province","country","lat
itude","longitude","geo_resolution","date_onset_s
ymptoms","date_admission_hospital","date_confirma
tion","symptoms","lives_in_wuhan","travel_history
_dates","travel_history_location","reported_marke
t_exposure","additional_information","chronic_dis
ease_binary","chronic_disease","source","sequence
_available","outcome","date_death_or_discharge","
notes_for_discussion","location","admin3","admin2
","admin1","country_new","admin_id"="15-
88"",1.0,1.0
"attribute_0=2 ==>
"id","age","sex","city","province","country","lat
itude","longitude","geo_resolution","date_onset_s
ymptoms","date_admission_hospital","date_confirma
tion","symptoms","lives_in_wuhan","travel_history
_dates","travel_history_location","reported_marke
t_exposure","additional_information","chronic_dis
ease_binary","chronic_disease","source","sequence
_available","outcome","date_death_or_discharge","
notes_for_discussion","location","admin3","admin2
","admin1","country_new","admin_id"="15-
88"",1.0,1.0
```

Fig. 6. Illustrative association rule.

```
Pattern,#SUP:,#CONF:
Deaths=0 ==> Country/Region='Mainland
China',777.0,0.128877094045447
Country/Region='Mainland China' ==>
Deaths=0,777.0,0.33448127421437795
Recovered=0 ==> Country/Region='Mainland
China',252.0,0.04061895551257253
Country/Region='Mainland China' ==>
Recovered=0,252.0,0.10848041325871717
Deaths=1 ==> Country/Region='Mainland
China',439.0,0.22757905650596164
Country/Region='Mainland China' ==>
Deaths=1,439.0,0.1889797675419716
Deaths=2 ==> Country/Region='Mainland
China',285.0,0.31181619256017507
Country/Region='Mainland China' ==>
Deaths=2,285.0,0.12268618166164443
Deaths=3 ==> Country/Region='Mainland
China',212.0,0.4
Country/Region='Mainland China' ==>
Deaths=3,212.0,0.09126130004304778
Deaths=6 ==> Country/Region='Mainland
China',201.0,0.6017964071856288
Country/Region='Mainland China' ==>
Deaths=6,201.0,0.0865260439087387
Deaths=0 ==>
Confirmed=1,1013.0,0.16802123071819539
Confirmed=1 ==>
Deaths=0,1013.0,0.9882926829268293
```

Fig. 7. Most robust association rules according to the selected parameters.

DS4C [34] dataset consists of four different CSV data files (Case data, patient data, time-series data, and additional data). The case data depicts the infectious cases data for COVID-19. The patient-data depicts the epidemiological and route data of COVID-19 patients. The time-series data illustrates the status of the COVID-19 patients. Furthermore, additional data is related to districts, weather, and nation. In this paper, the patient-data file is only used to accomplish the objective of our research.

Patient-data contains 5,165 patient's reports. Each patient report includes various characteristics such as sex, age, country, province, city, infection case, infected by contact number, symptom onset date, contact data, released date, deceased date, and state. The state characteristic depicts the label for each patient report. Every report in the dataset is either labeled isolated, released, or deceased. The states compared to the number of infection cases in the dataset are shown in Fig. 8. In this dataset, the unbalanced states of the distribution samples include 2,158 isolated states, 2,929 released states, and 78 deceased states. The sample distribution for sex is shown in Fig. 9. The sample distribution of age is shown in Fig. 10. The number of infected males is neared to the number of infected females. Moreover, the most infected age is from 20 to 29 years among patients. The causes of COVID19 infections in South Korean patients have been investigated.

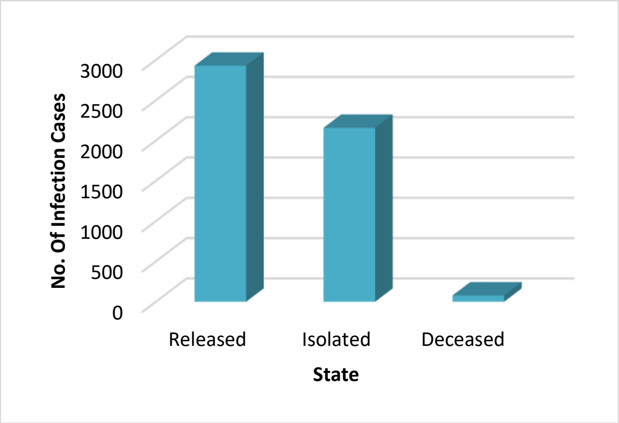


Fig. 8. The sample distribution for each state in the DS4C dataset.

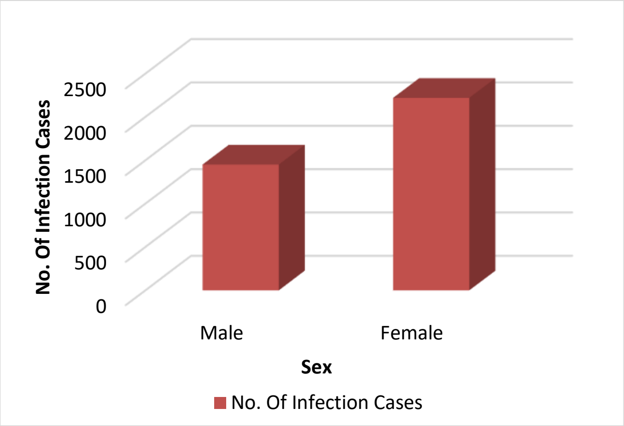


Fig. 9. The sample distribution for sex feature in DS4C dataset.

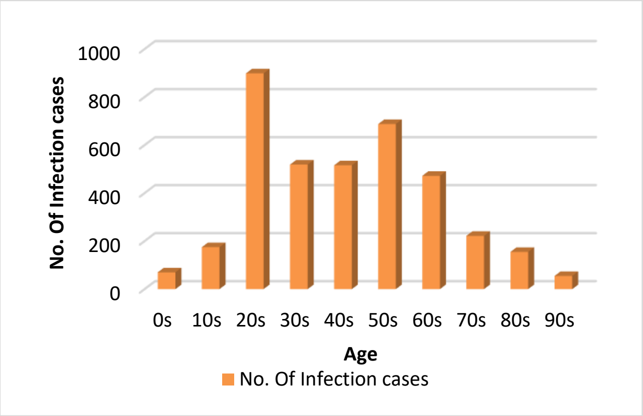


Fig. 10. The sample distribution of age features in the DS4C dataset.

V. Conclusion

COVID-19 is a virus that has spread worldwide. Data mining was instrumental in combating COVID-19. This paper aims to early diagnoses Corona Virus by using benchmark COVID-19 datasets by reprocess information provided for effective statistical analysis by using different data mining algorithms to identify meaningful patterns. The proposed algorithm (Apriori_Association_Rules) is developed by merging the hashed array mining technique and the Apriori transaction table to discover frequent associations. The Apriori and hashed array major advantages are that they eliminate the need for reconstructing conditional pattern-base, subtrees, and enhancing the tree construction operation. Extensive workflow tests of different types of data sets prove the effectiveness of the Apriori regarding the mining speed, memory utilization, and the number of discovered rules under various minimum support, and confidence levels. The results of the examination helped to better track, identify, and diagnose COVID-19. After all, the current research contributes to correct and reliable diagnosis and analysis of the Coronavirus disease.

Our future work is focusing on using artificial intelligence to extract primary features for timely and reliable detection of diseases in real-time applications.

References

[1] Anwar, H., & Khan, "Q. U. Pathology and Therapeutics of COVID-19: A Review", International Journal of Medical Students, 2020, doi: 10.5195/ijms.2020.498.

[2] H. Greenspan, B. van Ginneken and R. M. Summers, “Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique”, IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1153-1159,2016, doi: 10.1109/TMI.2016.2553401.

[3] M. J. Zaki, "Scalable algorithms for association mining", IEEE Transactions on Knowledge and Data Engineering, vol. 12, no. 3, pp. 372-390,2000, doi: 10.1109/69.846291.

[4] Webb, G. I., & Zhang, S., "Further Pruning for Efficient Association Rule Discovery", Lecture Notes in Computer Science, 2001, 2256, 605 - 618.

[5] Pinar Yazgana, Ali Osman Kusakci, "A Literature Survey on Association Rule Mining Algorithms", Southeast Europe Journal of Soft Computing, 2016, 5. 10.21533/sc journal.v5i1.102.

[6] Agrawal R, Mannila H, Srikant R, Toivonen H, Verkamo AI., "Fast discovery of association rules", Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R, editors, Advances in Knowledge Discovery and Data Mining. Menlo Park, CA: AAAI Press, 1996. p. 307-328.

[7] Huan Wu, Zhigang Lu, Lin Pan, Rongsheng Xu, and Wenbao Jiang, "An improved apriori based algorithm for association rules mining", IEEE Sixth International Conference on Fuzzy Systems and Knowledge Discovery, 2009, volume2, pages 51–55.

[8] Jiao Yabing, "Research of an improved apriori algorithm in data mining association rules", International Journal of Computer and Communication Engineering,2013, 2(1):25.

[9] Jiawei Han, Jian Pei, and Yiwen Yin, "Mining frequent patterns without candidate generation", ACM SIGMOD International Conference on Management of Data, 2001, pages 1–12.

[10] Grahne G and Zhu J, "Fast algorithms for frequent itemset mining using FP-trees," IEEE Transactions on Knowledge and Data Engineering,2005, vol. 17, no. 10, pp. 1347-1362, Oct. 2005. doi: 10.1109/TKDE.2005.166.

[11] Zhan, Foxiao&Zhu, Xiaolan& Zhang, Lei & Wang, Xuexi& Wang, Lu & Liu, Chaoyi, "Summary of Association Rules", IOP Conference Series: Earth and Environmental Science, 2019, doi: 10.1088/1755-1315/252/3/032219.

[12] Anil Vasoya, Nitin Koli, “Mining of Association Rules on Large Database Using Distributed and Parallel Computing,” Procedia Computer Science, vol. 79, pp. 221-230, 2016.

[13] J. Han, M. Kamber, "Data Mining: Concepts and Techniques, 3rd Edition", The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann,2011.

[14] S.C. Suh, "Practical Applications of Data Mining", Jones & Bartlett Publishers,2012.

[15] B.Venkatalakshmi, M.V Shivsankar, “Heart Disease Diagnosis Using Predictive Data Mining,” International Journal of Innovative Research in Science, Engineering and Technology, 2014.

[16] A Kate, Rohit & Nadig, Ramya., "Stage-Specific Predictive Models for Breast Cancer Survivability", International Journal of Medical Informatics, 2016, 97. 10.1016/j.ijmedinf.2016.11.001.

[17] A. Bellaachia and E. Guven, "Predicting breast cancer survivability using data mining techniques," Age, vol. 58, pp. 10-110, 2006.

[18] Nagesh Shukla, Markus Hagenbuchner, Khin Than Win, Jack Yang, "Breast cancer data analysis for survivability studies and prediction", Computer Methods and Programs in Biomedicine, 2018, Volume 155.

[19] W. Wu, H. Zhou, "Data-driven diagnosis of cervical cancer with support vector machine-based approaches", IEEE Access, 2017, vol. 5, pp. 25189–25195.

[20] Kate, Rohit, Nadig, Ramya, "Stage-Specific Predictive Models for Breast Cancer Survivability", International Journal of Medical Informatics, 2016, doi: 97. 10.1016/j.ijmedinf.2016.11.001.

[21] Jeffrey Dean, Sanjay Ghemawat, "Map-reduce: simplified data processing on large clusters", Conference on Symposium on operating systems Design and Implementation, 2004, pages10–10.

[22] Haoyuan Li, Yi Wang, Dong Zhang, Ming Zhang, and Edward Y. Chang.," PFP: parallel FP-growth for query recommendation, ACM Conference on Recommender Systems, 2008, pages 107–114.

[23] E. El-shafeyi, A. El-desouky, "A Big Data Framework for Mining Sensor Data Using Hadoop", Studies in Informatics and Control, 2017, ISSN 12201766, vol. 26(3), pp. 365-376.

[24] Hong. Jian Qiu, RongGu, Chunfeng Yuan, and Yihua Huang." YAFIM: A parallel frequent itemset mining algorithm with a spark", Parallel and Distributed Processing Symposium Workshops, 2014, pages1664–1671.

[25] Apache, Apache spark repository, 2016.

[26] Feng Zhang, Min Liu, Feng Gui, Weiming Shen, Abdallah Shami, and Yunlong Ma, "A distributed frequent itemset mining algorithm using spark for big data analytics", Cluster Computing, 2015, 18(4):1493–1501.

[27] X. Niu, M. Qian, C. Wu and A. Hou, "On a Parallel Spark Workflow for Frequent Itemset Mining Based on Array Prefix-Tree", IEEE/ACM Workflows in Support of Large-Scale Science (WORKS),2019, Denver, CO, USA, pp. 50-59.

[28] K. D. Rajab, "New Associative Classification Method Based on Rule Pruning for Classification of Datasets", IEEE Access, 2019, vol. 7, pp. 157783157795.

[29] Sornalakshmi, M., Balamurali, S., Venkatesulu, M. et al, "Hybrid method for mining rules based on enhanced Apriori algorithm with sequential minimal optimization in the healthcare industry", Neural Comput & Applic, 2020, doi.org/10.1007/s00521-020-04862-2.

[30] N. Karimtabar and M. J. Shayegan Fard, "An Extension of the Apriori Algorithm for Finding Frequent Items," 2020 6th International Conference on Web Research (ICWR), IEEE, 2020, pp. 330-334, doi: 10.1109/ICWR49608.2020.9122282.

[31] Sitarski, Edward Algorithm Alley, "HATs: Hashed array trees. ", Dr. Dobb's Journal, 1996; 21 (11). http://www. ddj.com/architect/184409965?pgno=5.

[32] https://github.com/beoutbreakprepared/nCoV2019/tree/master/covid19/data.

[33] http://github.io/targeting2019-ncov/cov.

[34] https://www.kaggle.com/kimjihoo/coronavirusdataset.

[35] Eman Khashan, Ali Eldesouky, M.Fadel, Sally Elghamrawy, "A Big Data-Based Framework for Executing Complex Query Over COVID-19 Datasets (COVID-QF) ", 2020.arXiv:2005.12271.

Mai Shawkat received her bachelor's degree in the Department of Communications and Information Engineering, Faculty of Engineering, Mansoura University in 2010. She enrolled as a researcher at the Faculty of Engineering, Mansoura University, with the major of Computer Engineering and Control Systems Department in 2013. Her research interests include data mining and artificial intelligence.

DOI: http://dx.doi.org/10.24018/ejece.2021.5.2.304

Mahmoud M. Badawy received his M.Sc. degree and Ph.D. in Computer Engineering and Control Systems from Mansoura University, Egypt. Currently, he is an associate professor at the Computer and Systems Engineering Department, Faculty of Engineering, Mansoura University, Egypt. Department of Computer Science and Informatics, Taibah University, Saudi Arabia. His research interests include Computer networking, Distributed control systems, Database Management systems, mobile robots, Internet of Things, and cloud computing.

Ali I. Eldesoky received the M.A. and Ph.D. degrees from the University of Glasgow, USA. He is currently a Full Professor with the Computers Engineering and Systems Department, Faculty of Engineering, Mansoura University, Egypt. He is also a visiting part-time Professor with MET Academy. He also teaches in American and Mansoura universities and has taken over many positions of leadership and supervision of many scientific articles. He has published hundreds of articles in well-known international journals.

DOI: http://dx.doi.org/10.24018/ejece.2021.5.2.304