Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Methods 203 (2022) 511–522

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Fuzzy association analysis for identifying climatic and socio-demographic factors impacting the spread of COVID-19

Sujoy Chatterjee [a], Deepmala Chakrabarty [b], Anirban Mukhopadhyay [c,*]

[a] Department of Informatics, School of Computer Science, University of Petroleum and Energy Studies, Dehradun, India
[b] Prasanta Chandra Mahalanobis Mahavidyalaya, West Bengal State University, India
[c] Department of Computer Science and Engineering, University of Kalyani, India

ARTICLE INFO

Keywords:
COVID-19
Association rule mining
Fuzzy association rules
Climatic factors
Socio-demographic factors

ABSTRACT

Recently, the whole world witnessed the fatal outbreak of COVID-19 epidemic originating at Wuhan, Hubei province, China, during a mass gathering in a film festival. World Health Organization (WHO) has declared this COVID-19 as a pandemic due to its rapid spread across different countries within a few days. Several research works are being performed to understand the various influential factors responsible for spreading COVID. However, limited studies have been performed on how climatic and socio-demographic conditions may impact the spread of the virus. In this work, we aim to find the relationship of socio-demographic conditions, such as temperature, humidity, and population density of the regions, with the spread of COVID-19. The COVID data for different countries along with the social data are collected. For the experimental purpose, Fuzzy association rule mining is employed to infer the various relationships from the data. Moreover, to examine the seasonal effect, a streaming setting is also considered. The experimental results demonstrate various interesting insights to understand the impact of different factors on spreading COVID-19.

1. Introduction

Coronavirus Disease 2019 (COVID-19) is an acute respiratory disease caused by a highly virulent novel coronavirus strain, SARS-CoV-2, which is a single-stranded RNA virus [1]. Due to its highly contagious nature, it rapidly propagates from person to person causing a pandemic situation worldwide. Since the first appearance in late 2019, the ongoing pandemic of COVID-19 has resulted in approximately 25,00,000 confirmed cases and more than 1,70,000 deaths in over 200 countries worldwide (https://coronavirus.jhu.edu/). In India, it has already caused more than 940705 confirmed cases and almost 98678 deaths (https://www.mohfw.gov.in/). Most of the initial efforts in analyzing the spread and outcome of COVID-19 focus on setting complex mathematical models to pandemic data for predicting the spread and peak of the disease transmission [2]. These works have mainly used the data of the number of cases reported daily in different COVID-19 tracker websites. It may be noted that the spread and vulnerability of COVID-19 vary from one place to another and from one person to another. Therefore, it is expected that certain climatic and socio-demographic factors, which vary from place to place and person to person, are likely to have an immense effect on determining the outbreak intensity and outcomes of COVID-19 pandemic. However, no systematic effort has been reported in the literature that deals with this issue. The proposed study addresses the problem of identifying important factors that explain the spread and outcome of COVID-19 through data-driven association analysis.

Association analysis is a rule-based machine learning technique that is used to discover interesting associations between the variables or attributes of a large data set [3,4]. It has been successfully utilized in biological information processing [5] and disease outbreak prediction [6]. In this proposed study, we mainly consider the problem of identifying important climatic and socio-demographic factors responsible for the intensity of COVID-19 outbreak in a particular country, state, or region. The intensity can be measured in terms of the number of cumulative cases of the disease at a particular point of time. The climatic and socio-demographic factors, such as the average temperature of the region, humidity, and population density, are considered as the potential predictor variables. We develop customized association rule learning techniques based on the available pandemic data from various trackers of COVID-19 cases and publicly available patient details of India and other countries to infer possible associations among the above variables and their influence in predicting the intensity and outcomes of COVID-19. The extracted association rules can also be used for

* Corresponding author.
E-mail address: anirban@klyuniv.ac.in (A. Mukhopadhyay).

https://doi.org/10.1016/j.ymeth.2021.08.005
Received 6 April 2021; Received in revised form 13 July 2021; Accepted 18 August 2021
Available online 22 August 2021

predicting future COVID-19 outbreaks and patients' survival chances.

Traditional association rule mining methods like Apriori algorithm [3,4] deal with categorical datasets, where each attribute–value pair is considered as an item. However, the attributes of the COVID-19 datasets, except a few, are mostly numeric and continuous. A possible solution to make these datasets ready for traditional rule mining is to categorize the quantitative attributes by defining some intervals taking coarser granularity. For example, the temperature attribute is continuous. This can be categorized by selecting two thresholds l and h so that a temperature t falls in interval low if $t \leqslant l$, in interval medium, if $l < t \leqslant h$, and in interval high if $t > h$. However, it is difficult to determine these thresholds universally as it is very subjective and the accuracy of the obtained rules is very sensitive to these threshold values. Therefore the better alternative is to define these intervals as linguistic variables in terms of fuzzy sets, where different intervals may overlap in fuzzy context.

In this proposed study, the work is performed in a twofold manner. Initially, the COVID data is collected for various countries with diverse characteristics for a particular time period. Then the corresponding socio-demographic data are collected from these countries. Finally, these data are combined as the raw dataset for our experiments. The values of different parameters (like temperature, humidity) change due to the seasonal effect. Therefore to understand the behavior of different attributes, various membership functions are needed to be defined.

Most of the research conducted on the COVID data considers the dataset as static data or time-varying data. As the data about different death rates and recovery rates are generated each day, this situation is considered as a streaming one. There is minimal research that deals with the dynamic behavior of the COVID data. This dynamic behavior arises from different preventive actions like lockdown, vaccination drives taken by the Government. Hence the relationship between the different attributes can vary with time, the rules generated from them also vary. In this situation, as time changes, the different climatic conditions like temperature, humidity, etc., also change. Therefore, the effect of various seasonal changes can be captured if the streaming fashion is considered. In this method, the weighted average approach is used to update the rules in a streaming manner. While considering the streaming condition, the most persistent rules over different seasons are also observed. Moreover, the goal is to discard some of the rules which are no longer effective while the new attributes arrive with the seasonal changes. This work aims to develop customized association analysis techniques based on COVID-19 pandemic data for understanding the impact of various climatic and socio-demographic factors. The main contributions of the research are summarized below.

- Identifying interesting associations among climatic and socio-demographic factors responsible for the region-specific outbreak intensity.
- Predicting possible region-specific future outbreaks and understanding the fuzzy relationships between various attributes in the fuzzy context.
- Identifying the relationship of different rules over the different time-spans considering the streaming mode and recognizing the persistent rules across different time windows.
- Providing a model to be used as a customized tool that can study the static behavior as well as dynamic behavior of different socio-demographic conditions on COVID data.

The rest of the paper is organized as follows. Section 2 describes the state-of-the-art approaches dealing with various COVID data to understand the behavior of the virus spread. Section 3 depicts the proposed methodology. The experimental design and results are explained in Section 4. Section 5 concludes the article by providing some future directions.

2. Related work

A spectrum of research works is being carried out to combat the COVID outbreak and forecast the possible spread [6–8]. One important work is the prediction of the actual spread of the pandemic. To deal with this issue, a set of classifiers, namely, SVM, logistic regression, neural network-based models, and two variants of Bayesian Network classifiers have been applied over the dataset of patients collected from STEMI [9]. In [10], the authors used the ARIMA model for forecasting the outbreak in 15 countries. The COVID-19 data including cumulative number of cases, cumulative number of deaths and recovery cases of top 15 affected countries in April 2020 were considered and they tried to predict 30 days forecast of COVID-19 outbreak where their prediction showed really very scary outcomes for especially some European countries like Italy, Spain, and France. In [11], an objective-based approach was proposed for the prediction of the continuation of COVID-19 using live forecasting. In this work, the authors tried to anticipate the live forecasting of COVID-19 assuming the past pattern will continue in the future. Here Exponential smoothing models were adopted to predict the forecast of COVID-19 confirmed cases because the Exponential smoothing family provides really good forecast accuracy especially for short series. Among other studies, the modified SEIR model was also used in [4] to design a model for COVID-19 pandemic considering quarantine and treatment. Here, the authors also applied the particle swarm optimization (PSO) algorithm on the data of Hubei province for estimating parameters of the SEIR model.

Additionally, many research works introduced different kinds of mathematical models like SIR, SEIR models for prediction, and tested the performance of the models on different real data collected from different countries [12,13]. Several analytical approaches of the SIR models have been introduced in the literature. As the different countries follow different strategies to control the spread of the epidemic, therefore, the SIR-based model can be adopted with the different local assumptions specific to the countries. It has been noticed that the major success of SIR models depends on the context of the applications and adoption of proper assumptions [14,15]. Hence, a large number of variants of SIR model, namely, SIS (susceptible-infectious-susceptible), SIRD (susceptible-infected-recovered-deceased), MSIR (Maternally-derived-immunity-susceptible-infected-recovered), SEIR (Susceptible-exposed-infected-recovered), SEIS (Susceptible-exposed-infected-susceptible), etc. have been considered as the popular methods to predict the COVID-19 spread. Another advanced version of the SIR model, namely, SIR-d model has taken into account another two important characteristics, namely, vital dynamics and constant population [15].

Zhang et al. [16] proposed a segmented Poisson model by using the power-law and exponential law to study the COVID-19 outbreak in six major countries. In another study, a parsimonious model was proposed that identified the infected individuals and fixed various measures for the containment policy. Apart from this, different deep learning-based techniques like Long short Term Memory (LSTM) models and curve-fitting have also been proposed [17] for prediction of the month-wise COVID-19 cases. Here the impact of various measures like social isolation and lockdown duration during that time are considered. Meanwhile, using the epidemiological SIR model, Khrapov et. al. [18] developed a mathematical model for forecasting the epidemic development of COVID-19 in China. Another group of researchers extended the SEIR model for understanding the importance of testing and quarantine policy [19].

Another interesting research mentioned in [20] attempted to detect the possible outbreak of COVID-19 pandemic in India, employing linear regression, Multilayer perceptron and Vector autoregression method. Machine learning-based model by employing the power of cloud computing framework for predicting the growth of COVID-19 in countries worldwide was presented in [21]. Another study reported in [22] utilized the Support Vector Regression (SVR) model to foresee the spread of novel coronavirus along with the number of patients who

would recover. They also used Pearson's Correlation measure to find the correlation between coronavirus and different weather conditions like temperature, humidity, and wind. Similarly, another recent work finds the correlation among a large number of countries and socio-economic indicators to predict COVID-19 spread using various machine learning techniques [23]. In this work, the authors employed a univariate feature selection method to choose the most relevant features (i.e., indicators) and use ANOVA for this purpose. Thereafter, based on the spread of COVID, the countries were classified into four categories and different traditional classifiers are used to classify the countries. But the number of classes based on the COVID statistics may not always be four and also due to several other external effects, the covid cases can abruptly rise up or fall down. Thus it makes the classification task more difficult. In the same line, another recent research [24] finds the environmental effects on COVID spread in different regions of India and in New York city. In this method a number of statistical analysis is performed in order to understand the effect of different environmental factors like Temperature, Relative humidity and COVID cases per day. This work finds the pairwise correlation between any two features. Importantly, the impact of different preventive measures like lockdown, vaccinations, etc can impose some kind of dynamic nature in the system and identifying those patterns over different instances with their stability is important. However, as the dataset is considered as a static one, the complex relationships among a large number of environmental indicators in streaming situation cannot be identified in this work [24].

Although most of the research works are aligned toward predicting the growth or pointing out the final size of the spread, a very limited study has been performed on how to find the effect of different socio-demographic factors over the COVID-19 spread. In the present study, we try to understand the complex relationships of various socio-demographic attributes affecting COVID-19 spread. Different seasonal attributes like temperature, humidity etc., and the various measures (like lockdown, imposing various restrictions, and locating some containment zones, etc.) taken by the Government of the respective countries may change over time. Here, during this process, some old rules are faded away and new rules can be generated. So identifying the most stable rules across different time windows is important for the authority or decision maker to make proper decisions. However, finding this kind of relationships in streaming situation to infer the effect of persistent associations among various environmental factors on COVID data has not been studied in other research works. Therefore, in this work, in addition to understanding the complex relationships among various socio-demographic factors, the static and dynamic behaviors of the COVID data are also captured by considering the association rule mining in both the situations, viz., static rule mining and streaming rule mining. This is expected to help adopt various measures to prevent the spread of the epidemic and take requisite healthcare initiatives.

3. Proposed methodology

In normal association rule mining methods like Apriori algorithm, the items i.e., attribute–value pairs are not numeric and continuous. Therefore, the traditional association rule mining fails to quantify some intervals (like low, middle and high temperature) considering the coarser granularity. As an alternative means, the intervals can be termed as the linguistic variables considering fuzzy sets. Here overlapping between the different linguistic variables over fuzzy set can remove the discrepancy.

A fuzzy association rule [25] will then look like $(X \text{ is } A_X) \Rightarrow (Y \text{ is } B_Y)$, where the itemset $X = \{x_1, x_2, \ldots, x_p\}$ is the antecedent and the itemset $Y = \{y_1, y_2, \ldots, y_p\}$ is the consequent. $A_X = \{F_{x1}, F_{x2}, \ldots, F_{xp}\}$ and $B_Y = \{F_{y1}, F_{y2}, \ldots, F_{yq}\}$ consist of the linguistic variables defined by fuzzy membership functions for the corresponding items in X and Y, respectively. An example of fuzzy rule can be (Temperature is *high*) ⇒ (Intensity is *high*). We aim to develop efficient fuzzy association rule learning algorithms to discover interesting fuzzy association rules from the region-wise COVID-19 pandemic data, where the different regions are considered as the transactions.

As mentioned above, consider I be the set of itemsets and the itemsets $X, Y \subseteq I$. Again, suppose $X \Rightarrow Y$ be the rule and T be a set of transactions. The support quantifies how frequently the itemsets occur together in the database. The support of an itemset X with respect to T is defined as the proportion of transactions in the dataset that contains X. Mathematically, this support can be expressed as $Supp(X) = \frac{|\{X \in T\}|}{|T|}$. The quantification of confidence of $X \Rightarrow Y$ is the proportion of transactions in the dataset that holds the item X, in which item Y also occurs. Confidence of the rule $X \Rightarrow Y$ is defined as $Conf(X \Rightarrow Y) = \frac{Supp(X \cup Y)}{Supp(X)}$.

For the region-wise data, we primarily consider the association rules in which the consequent is the intensity of the outbreak in a region. Fuzzy rules to be obtained from this data explain the relationships among the different factors and the intensity of the COVID-19 outbreak in a particular region. The factors or variables in the antecedent of the rules will be identified as the most relevant factors responsible for the intensity of the outbreak.

3.1. Association rule mining in static setting

Different symbols are utilized to represent the different attributes as shown in Table 1. This COVID data, including the socio-demographic data are taken. Table 2 contains region-wise COVID-19 data and their corresponding socio-demographic information like temperature, humidity and population density. Fuzzy membership sets utilized to map the quantitative data into sets are demonstrated in Figs. 2–6. The ten steps for extracting the different important rules are provided below.

- **Step 1:** First, the quantitative values of the different attributes are transformed into fuzzy membership values. From Table 2, consider 903 in the first record of attribute D as an example and it is converted into fuzzy set using fuzzy membership function. The membership values of 903 are (0.2425|mid + 0.2575|large) as 903 lies in both "mid" and "large" classes (as demonstrated in Fig. 3). This step is repeated for all the attributes of the dataset.
- **Step 2:** After converting all the values into fuzzy sets, fuzzy normalization process is done by using the ratio of two components. The first component is "The membership value of attribute X in one of its fuzzy class" and the second term is "Sum of the membership values of attribute X in all of its fuzzy classes".

  Applying the ratio of the two terms to the first record in Table 3, we get that 0.2425/(0 + 0.2425 + 0.2575) is equal to 0.4850 and 0.2575/(0 + 0.2425 + 0.2574) is equal to 0.5150. So, after normalization (0.4850|mid + 0.5150|large) become the new membership values of 903. Normalization is done for all attributes by following this step. While Table 3 shows a set of raw membership values of attribute D, Table 4 shows the normalized equivalents.
- **Step 3:** After normalization, we add all the normalized values of the same fuzzy class. Then, the summation value is divided by the total

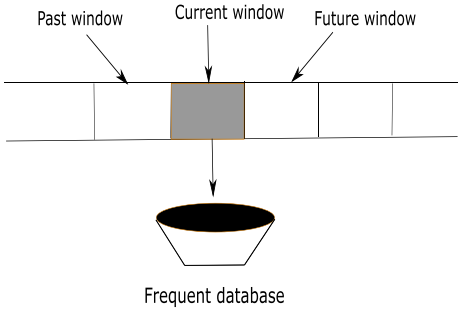

Fig. 1. Demonstration of streaming scenario.

Table 1
The symbols of the different attributes.

Parameters	Symbols
Temperature	T
Infected	I
Death	D
Recovery	R
Humidity	H
Population density	P

number of records present in the dataset to calculate the support of each fuzzy class and put them in set *Y*, where *Y* is called "itemsets". We consider attribute D as an example and the support values are determined individually for "small", "mid" and "large" fuzzy classes of attribute D as they are three different fuzzy classes of D. To demonstrate the calculation, here we choose "small" class of attribute D. We take the values of D.small from each record in Table 4 and add them together, i.e., $(0+0+0+1+0+1+1+1+1+1)=6$. After that, the summation value (here 6) is divided by the number of records (here 10) for obtaining the support of the "small" class of attribute D. So, here $(6/10)=0.6$ becomes the support of D.small.

- **Step 4:** Next, we compare each fuzzy class's support with the predefined minimum support value, which is 0.2 here. Taking the support of D.small from Table 4 as $0.6 > 0.2$, D.small is qualified as "frequent 1-itemset (Y1)" for further calculation. However, if the support of any fuzzy class (1-itemset) is smaller than the minimum support value, then that fuzzy class is rejected for further processing.
- **Step 5:** Fuzzy classes from frequent 1-itemset (Y1) are used to form all possible 2-itemset combinations. The support of each 2-itemset is calculated by selecting the minimum normalized value from two fuzzy classes (forming a 2-itemset) in each record. We sum up all the minimum values for a 2-itemset taken from all the records in the dataset, and then the summation value is divided by the total number of records. Table 5 contains the normalized values of two fuzzy classes {I.small} and {D.small} for 10 records. To demonstrate, consider {I.small, D.small} as a 2-itemset. In Table 5, for the first record {I.small} is 0 and {D.small} is also 0. Therefore, we get 0 for the first record.

  Similarly, we find the minimum values for 2–10 records. After that all of them are summed up, i.e., $(0+0+0+1+0+1+1+0+0+1)=4$ and then $(4/10)=0.4$ becomes the support value of 2-itemset {I.small, D.small}.
- **Step 6:** After getting the support of each 2-itemset, the support values are compared with the predefined minimum support value. If the support value of a 2-itemset is greater than or equal to the minimum support value, then the 2-itemset qualifies as "frequent 2-itemset (Y2)" for further calculation. Otherwise, it is rejected.
- **Step 7:** Forming all possible higher level combinations of itemsets, steps 5 and 6 are repeatedly performed until the point when there are no more combinations available. In this experiment, 3-itemsets and 4-itemsets are found to be developed.
- **Step 8:** From each Yi (i >=2), all possible association rules are extracted and then find the confidence value for each rule. Take if {P. low, H.wet} then {D.small} as an example and calculate the confidence value as $\frac{Support(P.low,H.wet,D.small)*\ 100}{Support(P.low,H.wet)} = \frac{0.2026\ *\ 100}{0.2893} = 70.03\%$.
- **Step 9:** After obtaining the confidence value of each rule, it is compared with the minimum confidence value, which was defined

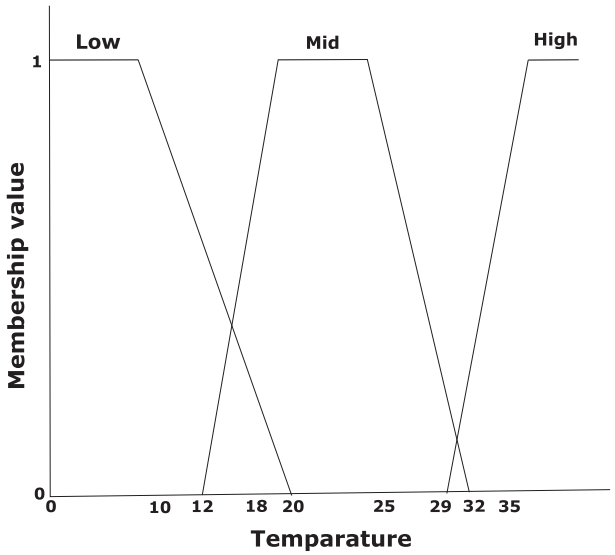

Fig. 2. Fuzzy membership sets for the temperature attribute.

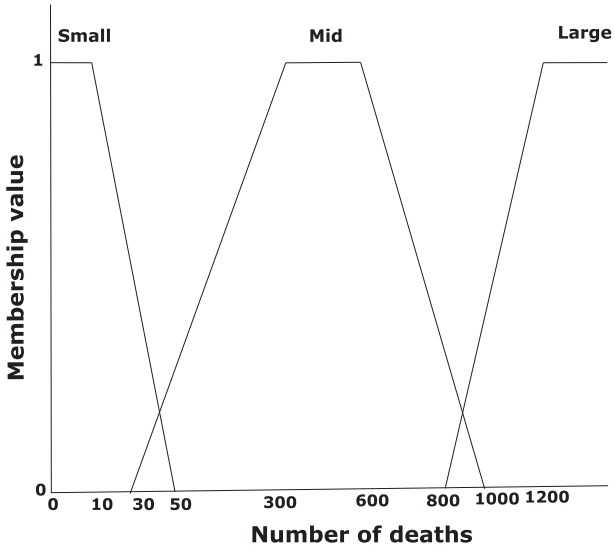

Fig. 3. Fuzzy membership sets for the Number of Deaths attribute.

Table 2
The historical data of COVID-19 pandemic (January 2020 to June 2020).

Country	Region	Total Infected after 50 Days (I)	Total Deaths after 50 Days (D)	Avg Temp (T)	Population density (P)	Avg humidity (H)
USA	Colorado	17364	903	5	19.9	61
USA	Wisconsin	7964	339	4	40.6	74
USA	Connecticut	25997	2012	7	285	61
USA	Alaska	371	10	0	0.49	75
USA	New York	257216	15302	6	159	58
INDIA	Kerala	28	0	29	859	71
INDIA	Telengana	873	23	30	312	59
INDIA	Rajasthan	1890	27	23	201	47
INDIA	Uttar Pradesh	1449	21	23	828	63
INDIA	Haryana	262	3	22	573	65

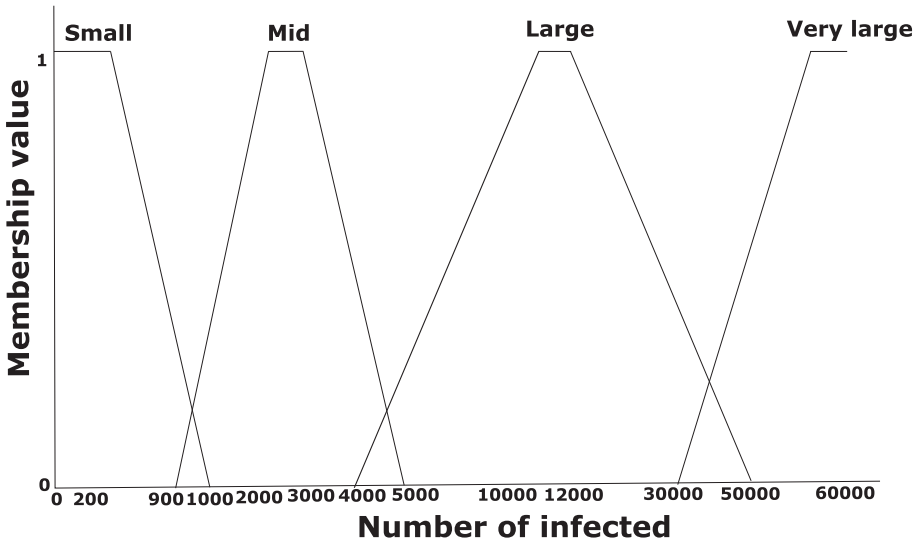

Fig. 4. Fuzzy membership sets for the Number of Infected attribute.

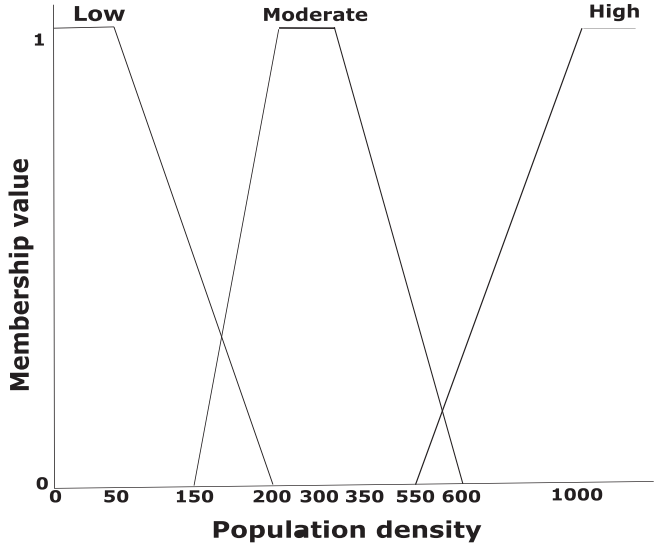

Fig. 5. Fuzzy membership sets for the population density attribute.

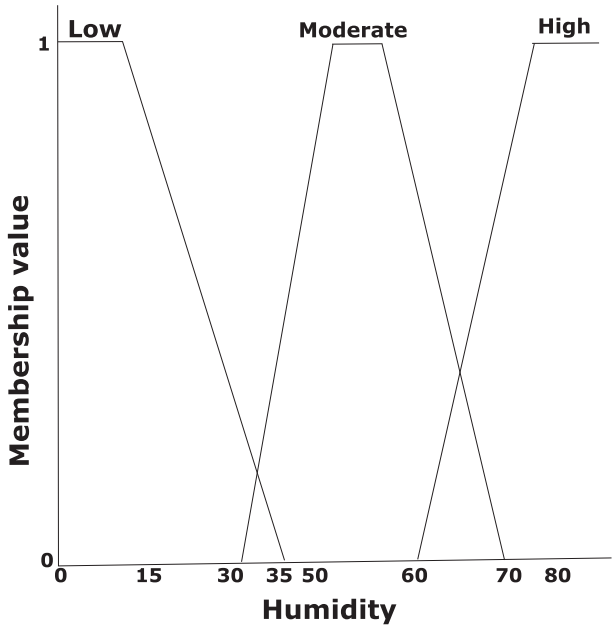

Fig. 6. Fuzzy membership sets for the humidity attribute.

earlier. In this experiment, we set the minimum confidence value to 60%. Therefore, only those association rules are reliable and qualified whose confidence values are greater than or equal to 60%. Otherwise, the rules are rejected.

- **Step 10:** Finally, those rules are treated as importance rules where D and I are present in the consequent part only because the effect of different attributes on death or infection can be realized better if it is present in the rule.

3.2. Association rule mining in streaming scenario

The proposed method for the streaming setting is described in a nutshell here. In this scenario, the data concerning COVID statistics arrive continuously. That means the number of deaths, infected and recovered persons are gathered in a day-by-day manner. The COVID statistics like temperature, population density and humidity may differ from region to region for a particular country. Hence, to study the behavior, we focus on a specific region and the day-wise data of COVID patients, including the climatic and socio-demographic data are collected. In this context, motivated by the work in [26], we introduce the streaming model for the COVID data considering the Fuzzy scenario, although the previous work does not consider the Fuzzy setting. In this work, a particular chunk of data, termed as window, is considered at a

Table 3
The raw membership values of attribute D.

Quantitative value of attribute D	D.small	D.mid	D.large
903	0	0.2425	0.2575
339	0	1	0
2012	0	0	1
10	1	0	0
15302	0	0	1
0	1	0	0
23	0.6750	0	0
27	0.5750	0	0
21	0.7250	0	0
3	1	0	0

Table 4
Normalized membership values for attribute D.

Quantitative value of attribute D	D.small	D.mid	D.large
903	0	0.4850	0.5150
339	0	1	0
2012	0	0	1
10	1	0	0
15302	0	0	1
0	1	0	0
23	1	0	0
27	1	0	0
21	1	0	0
3	1	0	0

Table 5
Normalized values of the two fuzzy class of I and D.

I.small	D.small
0	0
0	0
0	0
1	1
0	0
1	1
1	1
0	1
0	1
1	1

time. Therefore, in each time point, the frequent item-set algorithm is applied iteratively. In this context, we update the frequent item database every time. This frequent item database contains those itemsets that are present for a long time. At the same time, the support of each item is calculated for each time window. Therefore, if any of the old item-sets appears again in the next time window, the support value is computed using the weighted average approach. Similarly, the updation in the frequent item database is also needed to be performed. The step-by-step approach of the proposed model is described in the next few subsections.

- **Fetching Data using Sliding window**
  When the new data appears at each time-instant, there is a requirement to identify the changes and the distribution of the attributes. Therefore, we need to detect the changes in the distribution and relationships among the attributes. To remove the rules that are no longer valid, and to include new rules, we need to update the rule database accordingly. Fig. 1 illustrates this technique showing the different time windows of the data. A window contains a chunk of data collected in a given time period, e.g., the data collected in the first 15-days are stored in the first window, the data collected in the second window are for the next 15-days (excluding the first day), and so on.
- **Finding Frequent Itemsets**
  Here, all the attributes of the dataset are numeric. So, to find the frequent itemsets in the current window, we just follow the Steps 1 to 7 as described before in the methodology of Association Rule Mining in Static Setting (mentioned in Section 3.1).
- **Updating Frequent Itemset database**
  After collecting the frequent itemsets (with support value greater than or equal to the minimum support) for the current window, we update the frequent itemset database (FID). FID contains the itemsets that are either frequent for the current window, or they are not frequent for the current window but have been frequent for a long time (for the number of windows that is greater than or equal to the predefined minimum value), with their corresponding "stored support value". We calculate the "stored support value" of each itemset in the FID with the help of a simple weighted average strategy. The equation of the weighted average technique for calculating the stored support value for an item (denoted as S^i_{FID}) at i^{th} time instant is given below.

$$S^i_{FID} = \frac{(n-1)}{n} S^{(i-1)}_{FID} + \frac{S_{current}}{n} \quad (1)$$

Here, $S^{(i-1)}_{FID}$ denotes the stored support value of the same itemset in the last FID. $S_{current}$ be the support value of a frequent itemset (FI) in the current window. Considering the minimum support α, we apply the weighted average technique for a new frequent itemset (FI) in the current window with support greater than or equal to the minimum support α. In this scenario, for a new itemset the first part i.e., $((n-1)/n) * S^{(i-1)}_{FID}$ becomes zero as it was not present in the FID before and the second part, i.e., $(S_{current}/n)$ will be stored in the FID as the "stored support value" of the FI. In this situation, for a new itemset that appears for the first time, to alleviate the full weightage of it in the current window (as it appears for the first time in the current window), the second term $(S_{current}/n)$ is computed. Thus, the support of a new item that appears just one time in a new current window cannot be highly weighted. At the same time, if an itemset is found to be infrequent all of a sudden but has been frequent for a long time, instead of immediate removal of that itemset from the FID, we decrease its "stored support value" in the FID by using the above described weighted average technique. For this situation, $(S_{current}/n)$ will be zero and only $((n-1)/n) * S^{(i-1)}_{FID}$ part will be stored in the FID as the "stored support value" of the itemset. Let the minimum number of occurrences of an item in FID be θ. If the total number of occurrences of an itemset becomes less than θ (starting from the window where it first occurred as FI), only that itemset is deleted from the FID.
- **Updating Association Rules**
  When we update the FID for each window, simultaneously, the association rules extracted from the corresponding FID itemsets are kept updated. For each FID, all the possible association rules along with their corresponding confidence values are generated. Then, each rule's confidence value is compared with the predefined minimum confidence value (Let, β), and only the rules having confidence greater than or equal to β are kept.
- **Extracting Stable Association Rules**
  In the final step, the stable rules that are consistently present in different time windows are identified. Considering a stability threshold value as γ, if the frequency of obtaining the same rule from different windows is greater than or equal to γ, then the rule is considered stable. Thus, it signifies that even though some rules are not generated in successive windows, they are not removed immediately. Instead, the rules remain persistent in the database for a certain time, and thus it reflects the long-term dependency among the attributes showing the utility of the streaming approach.

4. Experimental design and results

In this section, the two different kinds of datasets used for the experiments (static and streaming scenario) are described. The experiment was performed in MATLAB 2015 and the environment is an Intel(R) CPU 2.4 GHz machine with 8 GB RAM running Windows 10.

4.1. Dataset preparation for static fuzzy association rule mining

World-wide 13 countries are selected to prepare the dataset for mining static association rule. These countries are USA, INDIA, ITALY, PAKISTAN, FRANCE, RUSSIA, NIGERIA, CHINA, JAPAN, AUSTRALIA, BRAZIL, SPAIN and SWEDEN. For each of these 13 countries, we further chose 3 to 5 specific regions (in total 83 regions) to collect region-wise COVID-19 pandemic data like the total number of infected individuals,

the number of persons who died after 50 days of the first instance of infection. The COVID-19 data for India is available in the link www.covid19india.org and the data for other countries are collected from the Johns Hopkins University tracker (https://coronavirus.jhu.edu/). To discover the relevant factors that could possibly explain the outbreak of COVID-19 in a particular region, we collect climatic attributes like Temperature, Humidity and socio-demographic data like Population-Density to be augmented in the raw data sets, whatever is available publicly. Here, all the attributes of the dataset are available in numeric format. Therefore, an effort is made to infer the interesting fuzzy association rules from the region-wise data across different countries that can help better decision-making and develop new strategies to alleviate the pandemic.

4.2. Dataset preparation for streaming fuzzy association rule mining

To prepare the data for the analysis considering the streaming scenario, we collected date-wise COVID-19 pandemic data like the cumulative number of infected, the cumulative number of persons recovered and the cumulative number of persons died from March 2020 to June 2020 for a particular region of India. In this respect, we choose one state of India, namely, 'Maharashtra'. The climatic data like 'average Temperature' and 'average Humidity' of the corresponding months and socio-demographic data like 'Population-Density' of Maharashtra is collected from the Internet.

4.3. Experimental results for static fuzzy association rule mining

In this experimental setting, we apply Fuzzy Association Rule Mining Technique for extracting different important rules connecting the diverse climatic and socio-demographic factors with COVID-19 pandemic data. For this experiment, we use 0.2 as the minimum support and 60% as the minimum confidence. Keeping the rules with the confidence value greater than or equal to 60%, some subsets of interesting rules are generated and those are reported in Table 6. Among the various rules evolved from the experimental analysis, some important rules are demonstrated in this table. It can be noticed that in the regions where the temperatures are medium, the number of deaths is small. Similarly, there is an effect of humidity over the number of deaths. It can be observed that if the humidity is wet, then the number of deaths also appears to be small.

Although the association between a small number of attributes can be perceived straightforwardly, it becomes difficult to understand the relationship between the different attributes for complex scenarios where multiple attributes exist. To exemplify this scenario, in the presence of multiple attributes like population density, humidity, temperature, etc., it becomes challenging to infer the relationships among them and understand the combined effect of different attributes over the number of deaths, number of infected people, or the number of recovered people.

4.4. Experimental results for streaming fuzzy association rule mining

In this experiment, we consider the sliding window size as 15. The minimum support value is considered as 0.5. Each time when the sliding window moves, the Frequent Itemset Database (FID) is updated accordingly. FID containing some frequent Itemsets (FIs) and their corresponding stored support values for different time periods are reported in Tables 7–11. The itemset I.small has a support value 1 for the first sliding window (according to step 3 of Section 3.1 and this support value is termed as actual support value hereafter), and it is greater than the minimum support value, i.e., 0.5. Thus, to compute the weighted support, we need to consider two factors, namely, stored support value and actual support value (as mentioned in Eq. (1)). Now stored support value means the support of the already existing itemset present in FID. I.small becomes a frequent itemset (FI) for this sliding window and is inserted into the current FID after evaluating its actual support value using Eq. (1). Here, the window's size is 15, and as it is the first sliding window, there was no past FID. Therefore, $((15-1)/15) * S_{FID}^{(i-1)}$ contributes to zero and the support value of I.small is (1/15) or 0.0667, which is stored in the current FID as mentioned in Table 7. If the itemset is consistently found as FI for some time period, then the stored support values in the corresponding FIDs are gradually increased. However, these values never exceed 1, because the maximum range of actual support value of any FI is 1.

In Table 8, the itemsets are generated after a particular instant while FID = 30. It can be observed that the stored support value of I.small is equal to 0.8127 because I.small constantly became frequent from the first sliding window up to 30^{th} sliding window. In this approach, if an itemset had been frequent for a long time, but now it has become infrequent, we do not remove it immediately from the FID. Here, we set the limit as 60%. If the total number of occurrences of an FI becomes less than 60% of the number of windows, starting from the first window when it was found to be frequent, the FI will be deleted from the FID. In Table 9, the stored support value of I.small is decreased to 0.2259 as I.small is not frequent in the current sliding window, but its total number of occurrences as an FI is still greater than or equal to 60%. However, I.small is no longer present in Table 10. Now, in this instance, I.small is purged out from the current FID. Some interesting rules, having confidence values greater than or equal to 70%, generated from different windows, are shown in Tables 12–14. Here, we consider 0.5 (50%) as the stability factor. If the number of times the same rule is obtained from different windows and the count is greater than or equal to 50% of the total number of windows, then the rule will be considered as a stable rule. Some interesting stable rules corresponding to their stability percentage evolved over the different instances (with different window sizes) are reported in Tables 15–17, keeping the rules with "Output Attributes" (D and R in this case) in the consequent part only. We also performed the experiments by varying the window size = 10 and the interesting frequent rules obtained for window numbers of 50 and 100 are demonstrated in Tables 18–19. In these experiments, we choose 50% as the threshold value for selecting the stable rules across different

Table 6
Some Interesting Rules generated by Static Association Rule Mining.

Rule	Antecedent	Consequent	Confidence
1	{T.mid}	D.small	78.11
2	{P.low}	I.small	60.90
3	{P.low}	D.small	70.31
4	{H.wet}	D.small	68.59
5	{T.mid, P.moderate}	D.small	76.35
6	{T.mid, H.wet}	I.small	62.10
7	{T.mid, H.wet}	D.small	75.95
8	{P.low, H.wet}	D.small	70.03
9	{P.low}	{I.small, D.small}	60.53
10	{T.mid, H.wet}	{I.small, D.small}	62.10

Table 7
Itemsets generated after the experiment while the size of sliding window is 15 and FID = 1.

Items	Support value
I.small	0.0667
D.small	0.0667
R.small	0.0667
T.mid	0.0667
H.wet	0.0667
P.moderate	0.0667
I.small, D.small	0.0667
I.small, R.small	0.0667
I.small, T.mid	0.0667
I.small, H.wet	0.0667

Table 8
Itemsets generated after the experiment while the size of sliding window is 15 and FID = 30.

Items	Support value
I.small	0.8127
D.small	0.8579
R.small	0.8738
T.mid	0.8738
H.wet	0.8738
P.moderate	0.8738
I.small, D.small	0.8127
I.small, R.small	0.8127
I.small, T.mid	0.8127
I.small, H.wet	0.8127

Table 9
Itemsets generated after the experiment while the size of sliding window is 15 and FID = 50.

Items	Support value
I.small	0.2259
I.mid	0.5544
D.small	0.3228
D.mid	0.5196
R.small	0.3564
T.mid	0.6200
H.wet	0.9682
P.moderate	0.9682
I.small, D.small	0.2259
I.small, R.small	0.2259
I.small, T.mid	0.2259
I.small, H.wet	0.2259

Table 10
Itemsets generated after the experiment while the size of sliding window is 15 and FID = 61.

Items	Support value
I.mid	0.2423
I.large	0.4860
D.mid	0.7393
R.small	0.1557
R.mid	0.7804
T.mid	0.2709
T.high	0.4560
H.wet	0.9861
P.moderate	0.9861
I.mid, H.wet	0.2423
I.mid, P.moderate	0.2423

Table 11
Itemsets generated after the experiment while the size of sliding window is 15 and FID = 70.

Items	Support value
I.large	0.7040
D.mid	0.5436
D.large	0.1456
R.mid	0.8529
T.mid	0.1560
T.high	0.5416
H.wet	0.9920
P.moderate	0.9920
I.large, H.wet	0.7040
I.large, P.moderate	0.7040

windows. However, it can be customized and any threshold value can be selected as per the requirement of the decision maker. To exemplify, in Table 15 and 16, as we fixed the threshold value as 50%, all the rules having greater value than 50% are retrieved. Here, by the nature of the

Table 12
Frequent rules obtained when the window size is 15 and window number is 35.

Rule	Antecedent	Consequent	Confidence
1.	I.small	D.small	100
2.	I.small	R.small	100
3.	T.mid	D.small	90.93
4.	{I.small, T.mid}	D.small	100
5.	H.wet	{D.small, R.small}	89.83
6.	{T.mid, H.wet}	D.small	90.93
7.	{I.small, H.wet, P.moderate}	{D.small, R.small}	100

Table 13
Frequent rules obtained when the window size is 15 and window number is 50.

Rule	Antecedent	Consequent	Confidence
1.	I.mid	D.mid	74.27
2.	I.mid	R.mid	70.26
3.	{I.small, T.mid}	R.small	100
4.	{I.mid, H.wet}	D.mid	74.27
5.	{I.mid, P.moderate}	R.mid	70.26
6.	{T.high, H.wet, P.moderate}	{D.mid, R.mid}	100

Table 14
Frequent rules obtained when the window size is 15 and window number is 85.

Rule	Antecedent	Consequent	Confidence
1.	H.wet	I.large	78.55
2.	I.large	D.large	70.99
3.	{I.large, T.mid}	R.large	80.54
4.	{I.large, T.mid, P.moderate}	{D.large, R.large}	80.54
5.	{T.mid, H.wet, P.moderate}	{D.large, R.large}	84.62
6.	{I.large, T.mid, H.wet, P.moderate}	{D.large, R.large}	80.54

Table 15
Some interesting stable rules when window size is considered as 15.

Rule	Antecedent	Consequent	Stabilty
1	{I.small, T.mid}	D.small	55%
2	{I.small, H.wet}	D.small	55%
3	{I.small, P.moderate}	D.small	55%
4	{I.small, T.mid}	R.small	55%
5	{I.small, H.wet}	R.small	55%
6	{I.small, P.moderate}	R.small	55%
7	{I.small, T.mid}	{D.small, R.small}	55%
8	{I.small, H.wet}	{D.small, R.small}	55%
9	{I.small, P.moderate}	{D.small, R.small}	55%
10	{I.small, T.mid, H.wet}	D.small	55%
11	{I.small, T.mid, P.moderate}	D.small	55%
12	{I.small, H.wet, P.moderate}	D.small	55%
13	{I.small, T.mid, H.wet}	R.small	55%
14	{I.small, T.mid, P.moderate}	R.small	55%
15	{I.small, H.wet, P.moderate}	R.small	55%
16	{I.small, H.wet, T.mid}	{D.small, R.small}	55%
17	{I.small, T.mid, P.moderate}	{D.small, R.small}	55%
18	{I.small, H.wet, P.moderate}	{D.small, R.small}	55%
19	{I.small,T.mid, H.wet, P.moderate}	D.small	55%
20	{I.small,T.mid,H.wet, P.moderate}	R.small	55%
21	{I.small, T.mid, H.wet, P.moderate}	{D.small, R.small}	55%

data, all these rules have equal stability values, i.e., 55% in Table 15. However, it may not always be the same. For example, different stability factors of the rules like 69% and 51% can be seen in Table 17. Besides this, if the decision maker decides to choose more stable rules then he/she may increase the level of threshold value and eventually it will generate more restrictive rules and the number of rules will become less. But these rules are more persistent throughout the whole period. In Table 18, it can be observed that $4^{th}, 5^{th}$ and the 6^{th} rules have the confidence 100% that denotes the importance of those rules. Similarly, in Table 19, the $1^{st}, 3^{rd}$ and 4^{th} rules have the confidence value 100%.

Table 16
Some interesting stable rules when window size is considered as 30.

Rule	Antecedent	Consequent	Stabilty
1	{I.small, T.mid}	D.small	50%
2	{I.small, H.wet}	D.small	50%
3	{I.small, P.moderate}	D.small	50%
4	{I.small, T.mid}	R.small	50%
5	{I.small, H.wet}	R.small	50%
6	{I.small, P.moderate}	R.small	50%
7	{I.small, T.mid}	{D.small, R.small}	50%
8	{I.small, H.wet}	{D.small, R.small}	50%
9	{I.small, P.moderate}	{D.small, R.small}	50%
10	{I.small, T.mid, H.wet}	D.small	50%
11	{I.small, T.mid, P.moderate}	D.small	50%
12	{I.small, H.wet, P.moderate}	D.small	50%
13	{I.small, T.mid, H.wet}	R.small	50%
14	{I.small, T.mid, P.moderate}	R.small	50%
15	{I.small, H.wet, P.moderate}	R.small	50%
16	{I.small, H.wet, T.mid}	{D.small, R.small}	50%
17	{I.small, T.mid, P.moderate}	{D.small, R.small}	50%
18	{I.small, H.wet, P.moderate}	{D.small, R.small}	50%
19	{I.small,T.mid, H.wet, P.moderate}	D.small	50%
20	{I.small, T.mid,H.wet,P.moderate}	R.small	50%
21	{I.small, T.mid, H.wet, P.moderate}	{D.small, R.small}	50%

Table 17
Some interesting stable rules when window size is considered as 10.

Rule	Antecedant	Consequent	Stability
1	{I.small, T.mid}	D.small	69%
2	{I.small, H.wet}	D.small	69%
3	{I.small, P.moderate}	D.small	69%
4	{I.small, T.mid}	R.small	69%
5	{I.small, H.wet}	R.small	69%
6	{I.small, P.moderate}	R.small	69%
7	{T.high, H.wet}	R.mid	51%
8	{T.high, P.moderate}	R.mid	51%
9	{I.small, T.mid}	{D.small, R.small}	69%
10	{I.small, H.wet}	{D.small, R.small}	69%
11	{I.small, P.moderate}	{D.small, R.small}	69%
12	{I.small, T.mid, H.wet}	D.small	69%
13	{I.small, T.mid, P.moderate}	D.small	69%
14	{I.small, H.wet, P.moderate}	D.small	69%
15	{I.small, T.mid, H.wet}	R.small	69%
16	{I.small, T.mid, P.moderate}	R.small	69%
17	{I.small, H.wet, P.moderate}	R.small	69%
18	{T.high, H.wet, P.moderate}	R.mid	51%
19	{I.small, H.wet, T.mid}	{D.small, R.small}	69%
20	{I.small, T.mid, P.moderate}	{D.small, R.small}	69%
21	{I.small, H.wet, P.moderate}	{D.small, R.small}	69%
22	{I.small,T.mid, H.wet, P.moderate}	D.small	69%
23	{I.small, T.mid, H.wet, P.moderate}	R.small	69%
24	{I.small, T.mid, H.wet, P.moderate}	{D.small, R.small}	69%

Table 18
Some frequent rules obtained when window size is 10 and window number is 50.

Rule	Antecedant	Consequent	Confidence
1.	{I.mid, H.wet	D.mid	78
2.	{I.mid, H.wet, P.moderate}	D.mid	78
3.	{I.mid, P.moderate}	R.mid	76
4.	{I.small, T.mid}	{D.small, R.small}	100
5.	{I.small, T,mid, P.moderate}	{D.small, R.small}	100
6.	{I.small, H.wet, P.moderate}	{D.small, R.small}	100

In a normal situation, the support values of an itemset for different time points are calculated. These support values change abruptly depending on the presence of the itemset in specific time windows. We consider a specific time-span to observe the behavior of the support value of a particular itemset temperature (i.e., T.mid) and notice the changes of the support value across different time durations as demonstrated in Fig. 7. The nature of the stored support value for the streaming

Table 19
Some frequent rules obtained when window size is 10 and window number is 100.

Rule	Antecedant	Consequent	Confidence
1.	{I.vlarge, H.wet, P.moderate}	{D.large, R.large}	100
2.	{I.large, H.wet}	D.large	82
3.	{I.large, T.mid, H.wet}	D.large	100
4.	{T.high, H.wet, P.moderate}	R.mid	100
5.	{I.large, H.wet, P.moderate}	D.large	82
6.	{T.mid. H.wet, P.moderate}	R.large	98

situation when the support values are weighted based on the previous window and current window is demonstrated in Fig. 8. Fig. 7 shows that the support value becomes constant for a specific time period and that it becomes zero during the window number 47 to 71. This is due to the absence of that item in the original data at that time-span. The immediate effect in support value is reflected while ignoring the long-term effect. However, from Fig. 8, the gradual rise and fall of the stored support value instead of abrupt changes can be noticed in the streaming situation. Similar observations can be made from Figs. 9 and 10, for the Infected number of people (i.e., I.small) in both the situations. To illustrate, the normal support value of another itemset I.small for different time windows is shown in Fig. 9 where sudden fall of support value is noticed after window number 25. Although, Fig. 10 exhibits the window-wise observation while considering the stored support value of I.small. Here, the stored support value of I.small gradually increases as long as this itemset is present across different time windows and its sudden absence in a particular time window did not remove this itemset from the database immediately, rather the stored support value decreases slowly.

In Table 6, from row number 10, we can observe the rule $\{T.mid, H.wet\} \Rightarrow \{I.small, D.small\}$. Basically, this type of situation happened in some states of India (e.g., Haryana) where the range of *T.mid* is between 12 and 32, while the value of *H.wet* is greater than 60. In this case, the threshold value of I.small and D.small are 1000 and 50, respectively. Similarly, it can be noticed from the 7^{th} rule that $\{T.mid, H.wet\} \Rightarrow D.small$ and this condition reflects the COVID situation in two provinces (e.g., Mayatte and Martinique) of France. Here the values of average temperature, humidity and number of deaths in Mayatte were 29, 77, and 11. Another important observation can be made from the 18th rule of Table 17. It can be inferred that the intensity of recovery becomes

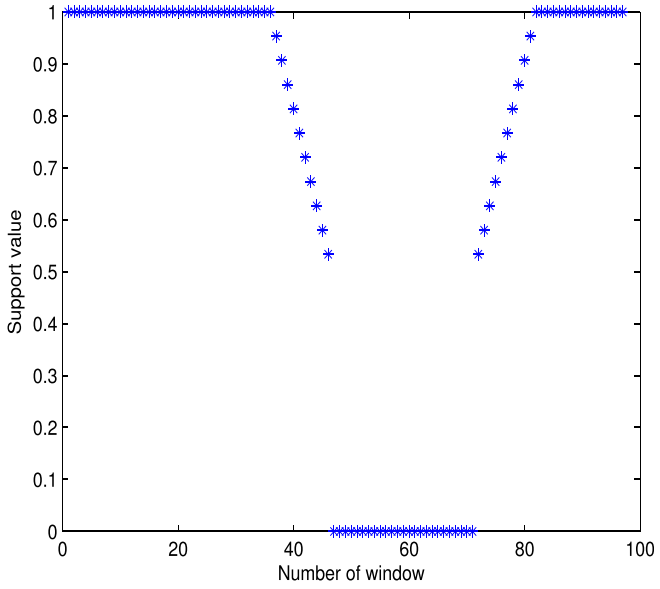

Fig. 7. Variation of support value for T.mid over different time windows while normal support value is considered.

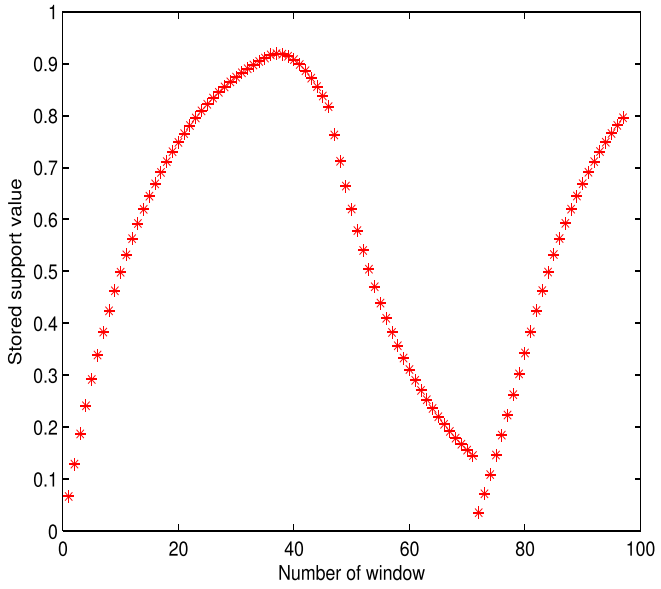

Fig. 8. Variation of stored support value for T.mid over different time windows in streaming situation considering weighted support value.

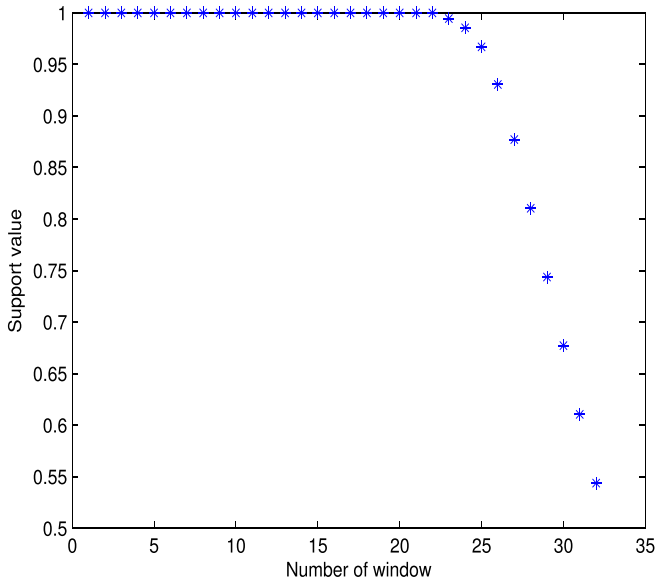

Fig. 9. Variation of support value for I.small over different time windows while normal support value is considered.

medium if the temperature, humidity and population density become high, wet and moderate, respectively. Interestingly, this rule is identified as the stable rule due to its persistence over the whole period of time considering all the time windows in streaming situation. As mentioned before, understanding these types of complex associations between the different intensity levels of climatic and socio-demographic factors become difficult. Thus, various preventive measures can be adopted based on the different outcomes observed from the derived fuzzy association rules. Importantly, as the streaming situation is also considered, policy makers may become aware of the seasonal effect on it. In this way, this proposed methodology can assist the decision makers and Government to choose the appropriate preventive measures to fight against COVID.

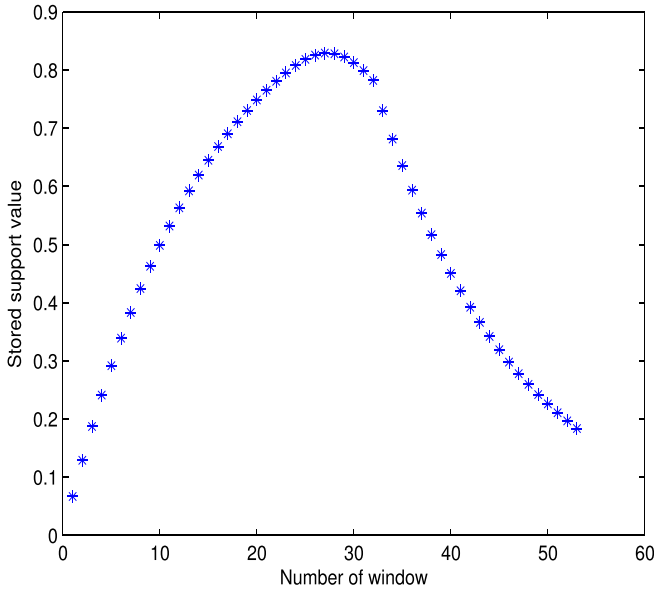

Fig. 10. Variation of stored support value for I.small over different time windows in streaming situation considering weighted support value.

4.5. Further analysis

In previous analysis for static situations as mentioned in Section 4.1, we consider the COVID statistics of the persons who died after 50 days of the first instance of infection. However, depending upon the nature of symptoms and incubation period, a COVID recovery can take a minimum of 14 days. Hence, we performed a further experiment on the COVID data of those 13 countries (comprising 81 regions) considering the number of persons who died after 14 days of the first instance of infection. The rules obtained from the experiment are demonstrated in Table 20. There are some interesting complex relationships that can be very difficult to understand without employing the proposed methodology. For example, T.mid and P.moderate jointly infer that the infection will become small. It is also seen T.low and H.wet implies that the death will be small. Furthermore, it can also be observed that if T.mid and P.moderate occur jointly then death will become small. So it can be noticed that similar kind of rules $\{T.mid, P.moderate\} \Rightarrow D.small$ and $\{P.low, H.wet\} \Rightarrow D.small$ also appear if 50 days of first instance of infection are taken into account (as mentioned in Table 6). However, the confidence of $\{T.mid, P.moderate\} \Rightarrow D.small$ is high in this situation.

To compare the performance with a recent work mentioned in [24], we conduct the same experiment on some common geographical locations e.g., New York City and Mumbai City considering the streaming case. In this experiment, we employed the same dataset available in [24]. The window size is kept fixed as 15. At first we carried out our analysis for New York city and some important rules extracted from this experiment with their corresponding stability percentage are shown in Table 21. In this first experiment, the temperature of New York city is within the range 5.17–27.5 degree Celsius. As shown in Fig. 2, the temperature attribute is divided into three fuzzy sets namely, T.high, T.

Table 20
Some interesting rules considering the incubation period of 14 days after the first instance of infection.

Rule	Antecedent	Consequent	Confidence
1	{T.low}	D.small	78
2	{T.low, H.wet}	D.small	82.2
3	{T.mid,P.moderate}	I.small	76.7
4	{T.mid,P.moderate}	D.small	92.2
5	{T.mid,H.wet}	{I.small,D.small}	74.1
6	{P.low,H.wet}	{I.small,D.small}	73.1

Table 21
Some interesting rules for New York City when window size is considered as 15.

Rule	Antecedent	Consequent	Stabilty
1	{T.low}	I.large	56.8%
2	{T.low, P.high}	I.large	56.8%
3	{T.high, P.high}	I.low	26.3%
4	{T.high,H.comfort,P.High}	I.low	26.3%
5	{T.high}	I.low	5%

Table 22
Some interesting stable rules for Mumbai City when window size is considered as 15.

Rule	Antecedent	Consequent	Stabilty
1	{T.high, P.high}	I.large	41.1%
2	{H.wet, P.high}	I.large	41.1%
3	{T.high,H.wet, P.high}	I.large	41.1%

mid and T.low. But here we consider the range of T.low is less than or equal to 12, whereas we set the range from 10 to 19 for T.mid and the range for T.high starts from 18 degree Celsius. From Table 21, one interesting stable rule can be noticed that exhibits infections become high when temperature becomes low and the stability is 56.8%. The threshold value of the stability factor is chosen as 0.5 (50%) like the previous experiment. It can be observed that another stable rule appears in the database which implies $\{T.low, P.high\} \Rightarrow I.large$. This similar kind of observation is also noticed in this work [24], where negative correlation exists between temperature and COVID cases per day. One more interesting rule reveals that infection becomes low when temperature remains high, humidity is within the comfort level and population density becomes high. As the stability of this rule is 26.3% (i.e., below 50% which is the threshold value as chosen), even though this is not a stable rule but it appears in the database. Thus the negative correlation between Temperature and COVID cases can be inferred from these rules for this New York City. We also performed another set of experiment on other geographical location i.e., Mumbai city and the corresponding rules obtained from the experiment are reported in Table 22. In this analysis, it can be observed that the joint effect of {T.high,H.wet,P.high} implies that Infection will be large. For this city also, similar kinds of characteristics can be observed in [24] i.e., infection will become higher when temperature becomes high. Hence, the positive correlation between temperature and COVID cases per day is also observed in this current experiment. Here the stability percentage values of all these important rules is 41.1 (e.g., less than 50%) while 40% is considered as threshold value. In this situation, the decision makers need to be less restrictive to choose the threshold value in order to obtain some associations. Although the streaming condition considering the dynamic behaviour of the data is not considered in this work [24]. Additionally, it is difficult to understand the joint effect of a large number of environmental characteristics for different instances of streaming situations if only the pairwise correlation is computed. Therefore, it is inconvenient in finding the most stable relationships/rules throughout the whole period. Thus, it demonstrates the utility and effectiveness of the proposed research over the socio-demographic data to adopt appropriate decisions.

5. Conclusion

Over the last many months, the whole world has been affected in different ways due to the fatal outbreak of COVID-19. Many research works have been performed during the pandemic to fight the rapid spread of COVID. Most of the research works are focused on forecasting the peak of the affected number of people using the improved SIR/SIER models. Some researches are also found to be very interesting to ease out the different decision making processes by predicting the size of the infections while considering the various control measures like incubation period (time lag between the infection and starting day of disease symptoms), latent period (time lag between infection to infectiousness), etc. However, limited research has been concerning with the different territories' socio-demographic conditions to understand the behavior of the number of deaths, infections, recovered, etc., based on the COVID data. Moreover, for complex scenarios, where multiple attributes are present, it becomes tough to find the relationship among the attributes and find the impact of the critical attributes over the number of deaths. In this article, we study the different characteristics of socio-demographic conditions like humidity, temperature, population, etc., to find the relationship among them with the virus's spread. This study is made considering the Fuzzy environment and we have applied static Fuzzy association rule mining for different countries comprising diverse geographical nature like temperature, humidity, population density, etc. In addition to that, as the data for the spread of COVID-19 are generated continuously, and the data is dependent on time, the generated data is considered to fit into a streaming model. As the different socio-demographic conditions like temperature, humidity, etc., vary in different seasons, this can significantly affect the spread of the virus. To understand this, we also developed the Fuzzy association rule mining for the streaming data and applied it to the COVID data using a time window-based model. The experimental results demonstrate the different interesting rules in both the conditions (static and streaming) that are expected to help the Government, policymakers, and healthcare persons implement different strategies to combat the epidemic. In this current study, we have considered region-wise COVID infection data. However, patient-wise data can be used to infer the possible outcome of survival of the patients. Additionally, considering the different measures like the number of lockdown days, the number of containment zones in a particular territory, etc., can also be considered to find the crucial relationships to cease the spread further.

This current study does not consider the patient-wise individual data during this period. However, the individual patient-wise data can be an interesting direction to understand the comorbidity along with the effect of different socio-economic conditions that can be studied in future. For example, a patient with high blood sugar and with chronic heart disease can be more prone to be infected. But it is observed that prediction of disease comorbidity can be erroneous if only the clinical data are employed. In addition to that, consideration of different attributes like population characteristics, age diversity, along with these socio-economic conditions can be helpful to understand the relationships between the diseases. To understand this type of analysis, mining of this kind of Fuzzy association rule can be one of the solutions. Interestingly, the analysis of different patient-wise statistics like age diversity among the number of cases of infected, deaths can be the important factors. More importantly, imposing the weight on the number of recoveries, infections and deaths are dependent on the period of lockdown, availability of vaccines in that locality. So two places which have an equal number of infections but having different days of lockdown period should be treated with different weights. As an example, the number of infections even after imposing ample days' of lockdown signifies the negligence of medical treatment by the local authority. Therefore, finding these kinds of relationships is also important to the authority in order to arrange various preventative measures. Another interesting direction in this work is considering the patient-wise transmission rate. As in this work we have not considered the patient-wise individual data, however, consideration of transmission rate along with age can infer other complex relationships. Thus, in this way, consideration on the period of lockdown for each country/state and study of a weighted Fuzzy Association rule can be more interesting to derive meaningful insights for adopting relevant strategies. The authors are working in these directions.

CRediT authorship contribution statement

Sujoy Chatterjee: Conceptualization, Data curation, Methodology, Writing - original draft. **Deepmala Chakrabarty:** Conceptualization, Data curation, Writing - original draft, Validation. **Anirban Mukhopadhyay:** Conceptualization, Investigation, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

A. Mukhopadhyay acknowledges the support received from the MATRICS project grant (MTR/2020/000326) of SERB, DST, Govt. of India.

References

[1] F.W., et al., A new coronavirus associated with human respiratory disease in china, Nature 579 (4) (2020) 1–8.
[2] A.J.K. et al., Early dynamics of transmission and control of covid-19: a mathematical modelling study, The Lancet Infectious Diseases 20.
[3] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, Proc. 20th Int. Conf. Very Large Data Bases VLDB 1215.
[4] S. He, Y. Peng, K. Sun, Seir modeling of the covid-19 and its dynamics, Nonlinear Dyn 101 (1667-1680).
[5] A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay, A novel biclustering approach to association rule mining for predicting hiv-1-human protein interactions, PLoS One 7 (4).
[6] A.L.B., et al., A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data, BMC Med. Inform. Decis. Mak. 12 (1) (2012) 124.
[7] C. Ceraolo, F. Giorgi, Genomic variance of the 2019-ncov coronavirus, J. Med. Virol. 92 (522–528) 2020.
[8] H. Wang, Z. Wang, e. a. Dong, Y., Phase-adjusted estimation of the number of coronavirus disease 2019 cases in wuhan, Cell Discov. 6 (462) 2020.
[9] X. Li, H. Liu, J. Yang, G. Xie, M. Xu, Y. Yang, Using Machine Learning Models to Predict In- ospital Mortality for ST-Elevation Myocardial Infarction Patients, in: Stud Health Technol Inform, vol. 245, 2017, pp. 476–480.
[10] P. Kumar, H. Kalita, S. Patairiya, Y.D. Sharma, C. Nanda, M. Rani, J. Rahmani, A.S. Bhagavathula, Forecasting the dynamics of covid-19 pandemic in top 15 countries in April 2020: Arima model with machine learning approach, medRxivdoi: 10.1101/2020.03.30.20046227.
[11] F. Petropoulos, S. Makridakis, Forecasting the novel coronavirus covid-19, PLOS One 15 (3) (2020) 1–8, https://doi.org/10.1371/journal.pone.0231236.
[12] R. Almeida, Analysis of a fractional seir model with treatment, Appl. Math. Lett. 84 (2018) 56–62.
[13] M. Osman, I. Adu, C. Yang, A simple seir mathematical model of malaria transmission, Asian Res. J. Math. 7 (2017) 1–22.
[14] B.F. Maier, D. Brockmann, Effective containment explains sub-exponential growth in confirmed cases of recent covid-19 outbreak in mainland china, medRxivdoi: 10.1101/2020.02.18.20024414.
[15] J.C. Miller, Mathematical models of sir disease spread with combined non-sexual and sexual transmission routes, Infect. Disease Modell. 2 (1) (2017) 35–55.
[16] X. Zhang, R. Ma, L. Wang, Predicting turning point, duration and attack rate of covid-19 outbreaks in major western countries. 135:109829.
[17] A. Tomar, N. Gupta, Prediction for the spread of covid-19 in india and effectiveness of preventive measures, Sci. Total Environ. 728 (2020), 138762.
[18] P.V. Khrapov, A.A. Loginova, Mathematical modelling of the dynamics of the coronavirus covid-19 epidemic development in china, Int. J. Open Inf. Technol. 8 (4).
[19] D. Berger, K. Herkenhoff, S. Mongey, An seir infectious disease model with testing and conditional quarantine, Working Paper 26901, National Bureau of Economic Research (March 2020). doi:10.3386/w26901.
[20] R. Sujath, J. Chatterjee, A.E. Hassanien, A machine learning forecasting model for covid-19 pandemic in india, Stoch. Env. Res. Risk Assess. 34 (2020) 959–972.
[21] S. Tuli, S. Tuli, R. Tuli, S.S. Gill, Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing, Internet Things 11 (2020), 100222, https://doi.org/10.1016/j.iot.2020.100222.
[22] M. Yadava, M. Perumal, M. Srinivas, Analysis on novel coronavirus (covid-19) using machine learning methods, Chaos, Solitons Fract. 139 (2020), 110050.
[23] A. Mihoub, H. Snoun, M. Krichen, M. Kahia, R. Bel, H. Salah, Predicting covid-19 spread level using socio-economic indicators and machine learning techniques, in: SMARTTECH 2020 – The First International Conference of Smart Systems and Emerging Technologies, Nov 2020, Riyadh, Saudi Arabia, 2020.
[24] H. Bherwani, A. Gupta, S. Anjum, A. Anshul, R. Kumar, Exploring dependence of covid-19 on environmental factors and spread prediction in india, npj Clim. Atmos. Sci. 3 (38) 2020.
[25] G. Ho, W. Ip, C. Wu, Y. Tse, Using a fuzzy association rule mining approach to identify the financial data association, Expert Syst. Appl. 39 (10) (2012) 9054–9063.
[26] R. Khade, J. Lin, N. Patel, Frequent set mining for streaming mixed and large data, in: 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), 2015, pp. 1130–1135, https://doi.org/10.1109/ICMLA.2015.218.