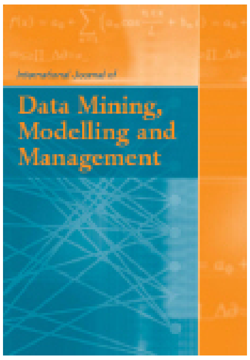


International Journal of Data Mining, Modelling and Management

ISSN online: 1759-1171 - ISSN print: 1759-1163
https://www.inderscience.com/ijdmmm

Application of rule-based data mining in extracting the rules from the number of patients and climatic factors in instantaneous to long-term spectrum

Sima Hadadian, Zahra Naji-Azimi, Nasser Motahari Farimani, Behrouz Minaei-Bidgoli

DOI: 10.1504/IJDMMM.2023.10055202

Article History:

Received:	13 October 2020
Last revised:	02 January 2021
Accepted:	02 July 2021
Published online:	04 April 2023

Application of rule-based data mining in extracting the rules from the number of patients and climatic factors in instantaneous to long-term spectrum

Sima Hadadian, Zahra Naji-Azimi* and Nasser Motahari Farimani

Department of Management,
Faculty of Economics and Administrative Sciences,
Ferdowsi University of Mashhad (FUM),
Mashhad, Iran
Email: sima_hadadian@yahoo.com
Email: znajiazimi@um.ac.ir
Email: n.motahari@um.ac.ir
*Corresponding author

Behrouz Minaei-Bidgoli

Computer Engineering School,
Iran University of Science and Technology,
Tehran, Iran
Email: b_minaei@iust.ac.ir

Abstract: Predicting the number of patients helps managers to allocate resources in hospitals efficiently. In this research, the relationship between the number of patients with the temperature, relative humidity, wind speed, air pressure, and air pollution in instantaneous, short-, medium- and long-term indices was investigated. Genetic algorithm and ID3 decision tree have been used for feature selection, and classification based on multidimensional association rule mining algorithm has been applied for rule mining. The data have been collected for 19 months from a pediatric hospital whose wards are nephrology, hematology, emergency, and PICU. The results show that in the long-term index, all climatic factors are correlated with the number of patients in all wards. Also, several if-then rules have been obtained, indicating the relationship between climate factors in four indices with the number of patients in each hospital ward. According to if-then rules, optimal planning can be done for resource allocation in the hospital.

Keywords: temperature; relative humidity; wind speed; air pressure; air pollution; patients; hospital; association rule mining; classification; genetic algorithm; ID3 decision tree.

Reference to this paper should be made as follows: Hadadian, S., Naji-Azimi, Z., Motahari Farimani, N. and Minaei-Bidgoli, B. (2023) ‘Application of rule-based data mining in extracting the rules from the number of patients and climatic factors in instantaneous to long-term spectrum’, *Int. J. Data Mining, Modelling and Management*, Vol. 15, No. 1, pp.30–52.

Biographical notes: Sima Hadadian is a researcher at Ferdowsi University of Mashhad (FUM), Mashhad, Iran. She has expertise in quantitative and statistical methods, data mining, fuzzy logic, and optimisation. She is also interested in business and healthcare problems

Zahra Naji-Azimi is an Associate Professor in Department of Management, Faculty of Economics and Administrative Sciences at Ferdowsi University of Mashhad (FUM), Mashhad, Iran. She has expertise in operational research, especially in combinatorial optimisation and heuristic and metaheuristic algorithms. She is also interested in healthcare problems and has some papers and researches in fuzzy and stochastic programming problems.

Nasser Motahari Farimani has PhD in Management from the University of Tehran. He is an Associate Professor of Industrial Management at the Ferdowsi University of Mashhad (FUM), Iran (2013-present). His areas of research include process engineering, information system, organisational improvement methods.

Behrouz Minaei-Bidgoli obtained his PhD from Michigan State University, Michigan, USA in Computer Science and Engineering Department. He is an Associate Professor in the School of Computer Engineering at the Iran University of Science and Technology, Tehran, Iran. He is leading a research group in data mining as well as another one in video game technologies. His research interests include text mining, natural language processing, and machine learning.

1 Introduction

It is of great importance to consider the relationship between various climatic factors and the number of patients to use human resources and medical facilities in hospitals optimally. Since the impact of climatic factors, especially air pollution, on human health varies from hours to days, dividing the effects of climatic factors from instantaneous to long-term indices provides a more detailed analysis of this relationship. Also, climatic factors affect individuals differently, and estimating the number of patients should not be considered the same for all individuals of different ages (Toti et al., 2016).

Investigating the relationship between climatic factors and the number of patients requires long-term data collection. Data mining is used to discover and extract information from databases. Among all data mining methods, rule mining is used to find out significant relationships between the set of items in large databases. Especially, association rule mining is one of the main data mining techniques, and of course, the most essential form of discovering and extracting patterns in learning systems that presents the relationships as if-then statements (Weng, 2016). Classification based on association rule mining, like association rule mining seeks to find the connections between items and values as if-then statements. But unlike association rule mining, in classification based on association rule mining, the ‘then’ section of rules only contains predefined values or items, which is called a ‘class’.

The issue addressed by this study is rules extraction to predict the number of patients based on climatic factors, that can be used for the optimal allocation of resources. The lack of a suitable system to check the balance between demand and the supply of hospital resources for different days of the year, or in other words, the lack of knowledge about the number of patients in the coming days causes inaccurate planning for the supply of hospital resources. This study aims to discover hidden knowledge between climatic factors such as temperature, relative humidity, wind speed, air pressure, and air pollution in the form of four indices: instantaneous index (the value of climatic factors in patient's referral day), short-term (the average of climatic factors in one week before the patient's referral), medium-term (the average of climatic factors in two weeks before patient's referral) and long-term (the average of climatic factors in one month before referral) with the number of patients in emergency, hematology, nephrology and PICU wards. This research focuses on a pediatric hospital and the reason for choosing such a hospital is their patients who are only children, as patients in this age range are more affected by climatic factors.

In this study, to discover the rules, genetic algorithm (GA) and ID3 decision tree (ID3) algorithm have been used for feature selection and classification based on multidimensional association (CBMA) rule mining have been used for rule mining. One of the innovations of this study is considering air pollution along with other climatic factors to classify and predict the number of patients. Another innovation of this research is the transformation of climatic factors into four indices: instantaneous, short-, medium- and long-term. Moreover, the method used in this research to predict the number of patients is another aspect of innovation.

In the following, in Section 2, we review previous studies. In Section 3, we describe the different methods used in this study. In Section 4, we introduce the case study, and in Section 5, we present the computational results. Finally, Sections 6 and 7 include discussion and conclusion.

2 Related works

By reviewing previous studies, it is observed that few studies have focused on predicting the number of patients. Toti et al. (2016) and Martín Martín and Sánchez Bayle (2018) predicted based on only one climatic factor. On the other hand, several studies have investigated the impact of climatic factors on pediatric diseases. Tauler et al. (1985) showed a significant relationship between air pollution and pediatric asthma using an analysis of variance. In addition, D'Souza et al. (2008) confirmed the relationship between temperature and humidity with pediatric viral diarrhea using the linear regression method. Mireku et al. (2009) demonstrated the effect of temperature, humidity, and air pressure on childhood asthma by time series analysis. Onozuka and Hashizume (2011) investigated the relationship between the temperature and humidity with infectious diseases in children using time series analysis. Hervás et al. (2015) examined the impact of different seasonal weather conditions on children having asthma with the use of the regression model. Kim et al. (2017) investigated the effect of temperature, pressure, and humidity on seizure and epilepsy in children using a nonlinear regression model. Martín Martín and Sánchez Bayle (2018) demonstrated the association between temperature and air contaminants with children's respiratory diseases by the use of multiple regression models. Moradiasl et al. (2018) showed a significant direct relationship between visceral leishmaniasis, which is the most commonly found disease among children, the hot temperature and sunny days. Yu et al. (2018) showed that pediatric epistaxis is related to the temperature and is not related to the humidity using a

Poisson regression model. Liu et al. (2019) investigated the effect of temperature and relative humidity on paramyxovirus respiratory syncytial virus, parainfluenza virus, and human metapneumovirus using multiple linear regression. Li et al. (2020) explored the correlation between the seasonality of pandemic Influenza and temperature, relative humidity, and PM1 concentrations through preliminary Pearson's r correlation test and subsequent time-series Poisson regression analysis using the distributed lag nonlinear model.

On the other hand, numerous studies have used association rule mining in the field of medicine. Ordonez et al. (2001) investigated the relationship between different factors and heart disease using association rule mining. By introducing a new association rule mining algorithm, Ordonez (2006) examined the relationship between age, gender, and smoking with heart diseases. Nahar et al. (2013) conducted a study to investigate the relationship between individuals' characteristics and heart diseases using association rule mining. Ivancevic et al. (2015) studied the relationship between children's characteristics and premature tooth decay using association rule mining. Toti et al. (2016) examined the relationship between air pollution and airborne elements with childhood asthma attacks by association rule mining. Borah and Nath (2018) investigated the relationship between factors affecting breast cancer and cardiovascular diseases using the FP-growth method, association rule mining, and the proposed Single Scan Pattern Tree algorithm. Alwidian et al. (2018) introduced a new association rule mining algorithm to predict breast cancer, considering influential factors. Wang e al. (2019) explored medical comorbidities such as sleep disorders and digestive diseases of mental disorders using association rule mining. Sarıyer and Öcal Taşar (2020) discovered frequent rules between diagnosis types and different types of laboratory diagnostic tests (LDTs) for emergency ward patients using association rule mining. Alaiad et al. (2020) predicted chronic kidney disease through effective factors such as age, blood pressure, red blood cells etc., using naive Bayes, decision tree, support vector machine, K-nearest neighbour, and a classifier based on association rule mining. Furthermore, they showed that a classifier based on association rule mining and k-NN achieved the highest accuracy.

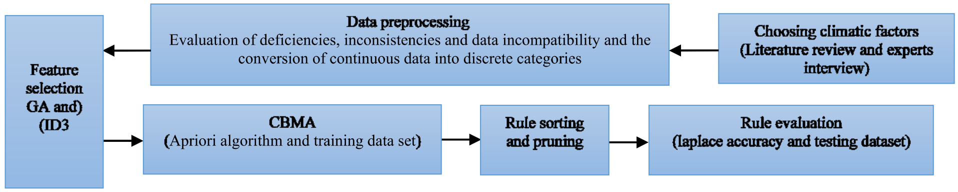
Previous studies have investigated the effect of different factors on one or some diseases. However, only a limited number of climatic factors have been taken into account in each research, and the impact of climatic factors has not been considered in instantaneous, short-, medium- and long-term indices. Also, previous studies have not provided a comprehensive insight into the number of patients of a hospital for better management strategies; therefore, the present study aims to address these shortcomings. We extract the rules between the number of patients and different climatic factors i.e., air temperature, relative humidity, wind speed, air pressure, and air pollution in instantaneous to long-term spectrum focusing on children patients, which has not been considered in any of the previous papers.

3 Research method

This is practical research to discover the rules based on the relationship between climatic factors such as temperature, relative humidity, wind speed, air pressure, and air pollution in instantaneous, short-, medium- and long-term indices with the number of patients in different wards of a pediatric hospital. For this purpose, data mining approaches have been applied. In particular, for feature selection (extracting influencing factors), GA and

ID3 have been used, and for rule mining, CBMA has been applied. The general stages of the research are presented in Figure 1.

Figure 1 The research methodology (see online version for colours)


Data preprocessing
Evaluation of deficiencies, inconsistencies and data incompatibility and the conversion of continuous data into discrete categories
Choosing climatic factors (Literature review and experts interview)
Feature selection GA and) (ID3
CBMA (Apriori algorithm and training data set)
Rule sorting and pruning
Rule evaluation (laplace accuracy and testing dataset)

3.1 *Climatic factors*

In this study, climatic factors including temperature, relative humidity, wind speed, air pressure, and air pollution were chosen by reviewing previous research and the opinion of the experts, such as pediatricians. By studying previous research, four climatic factors, including air temperature, relative humidity, air pressure, and air pollution were selected. These factors were confirmed by experts, and the wind speed factor was added to them, thus in total, five climatic factors have been selected in this study (the experts include six pediatricians of the studied hospital). Each climatic factor will be transformed into four indices: instantaneous, short-, medium- and long-term. Therefore, 20 climatic factors will be considered. Table 1 shows the effect of climatic factors in previous studies.

Table 1 Extraction of climatic factors from previous research

Climatic factors	*Type of disease*	*The impacts*	*Reference*
Temperature	Gastroenteritis	The decrease in temperature worsens the disease.	Tauler et al. (1985)
Air pollution	Asthma	The increase in air pollution worsens the disease.	D'Souza et al. (2008)
Temperature and relative humidity	Viral diarrhea	The increase in temperature and decrease in humidity worsen the disease.	Mireku et al. (2009)
Temperature, humidity and air pressure	Asthma	There is no correlation between air pressure and the disease, the increase in temperature and humidity worsens the disease.	Onozuka and Hashizume (2011)
Temperature and humidity	Infectious diseases	The increase in temperature and humidity worsens the disease.	Kim et al. (2017)
Temperature, humidity and air pressure	Seizures and epilepsy	There is no correlation between air pressure and humidity with the disease, the increase in temperature improves the disease.	Martín Martín and Sánchez Bayle (2018)

Table 1 Extraction of climatic factors from previous research (continued)

Climatic factors	*Type of disease*	*The impacts*	*Reference*
Temperature	Visceral leishmaniasis	The increase in temperature worsens the disease.	Moradiasl et al. (2018)
Temperature and humidity	pediatric epistaxis	There is no correlation between humidity and the disease, the increase in temperature improves the disease.	Yu et al. (2018)
Temperature and humidity	Paramyxovirus respiratory syncytial virus, parainfluenza virus, and human metapneumovirus	There is a negative correlation between paramyxovirus respiratory syncytial virus and human metapneumovirus with humidity, the increase in temperature worsens parainfluenza virus.	Liu et al. (2019)
Temperature and PM1 concentrations	Pandemic influenza	Low temperature and high PM1 worsen the disease.	Li et al. (2020)

3.2 Data preprocessing

Data preprocessing is the first step in data mining, which eliminates deficiencies and inconsistencies. Data preprocessing consists of different parts such as data cleaning, data integration, data reduction, and data transformation; each is used based on the needs and features of the available data (North, 2012; Zhang and Zhang, 2002). In this study, missing data are considered and data are transformed into discrete variables. In Section 5.1, data preprocessing will be completely explained.

3.3 Feature selection

In this study, based on Figure 1, after selecting climatic factors by the literature review and experts' interview and doing data preprocessing, feature selection is performed using a hybrid method (Khan and Baig, 2015). In various processing, it has been concluded that using all available features that appear to be related to the studied variable, does not necessarily lead to a high performance of extracting rules from the databases. Typically, if some of the related features are used, speed and accuracy would increase (Javed et al., 2012).

There are various methods for feature selection, all of which consist of two main functions: Generation function and Evaluation function. The generation function locates the candidates with different features, while the evaluation function evaluates the candidate feature subsets according to a specific method (Dash and Liu, 2003). In this research, because of the large number of feature subsets [which makes it NP-Hard to consider all of them (Chandrashekar and Sahin, 2014)], the GA is used in the generation function, and ID3 decision tree algorithm is applied in the evaluation function. A GA is a specific kind of evolutionary and nature-inspired algorithm that is used to find approximate solutions for optimisation problems. In the following, the feature selection process is described by the hybrid method proposed by Khan and Baig (2015) and illustrated in Figure 2.

3.3.1 The hybrid algorithm for feature selection

The hybrid algorithm consists of the GA and ID3 decision tree. In the first stage of the hybrid GA, some chromosomes are selected as the initial population in the search space, in which the length of each chromosome equals the number of the features, i.e. 20. Every gene can have a value of 0 or 1, in which 1 indicates the presence of the feature and 0 indicates the absence of the feature. In the second stage, the fitness function is calculated for each chromosome. The fitness function is calculated by the ID3 decision tree algorithm, and the accuracy is calculated by comparing the testing datasets with the tree output. The more accurate chromosome is considered the best. The ID3 decision tree uses the information gain formula (1), which is based on the concept of entropy, to construct the tree (Xiaohu et al., 2012; Lai et al., 2016).

Assume that a dataset D has n samples with k classes. Let $P(C_i, D)$ represents the proportion of C_i in D, where $C_i(i = 1,2,\ldots,k)$ are the set of samples that belong to the i^{th} class. The entropy of the dataset can be calculated by formula (1):

$$Entropy(D) = -\sum_{i=1}^{k} P(C_i, D) \times Log_2 P(C_i, D) \tag{1}$$

Let us assume that the feature A has different values in $v = \{v_1, v_2,\ldots,v_m\}$ and let D_j be those members of D in which $A = v_j$, the entropy of the dataset from feature A is given by formula (2):

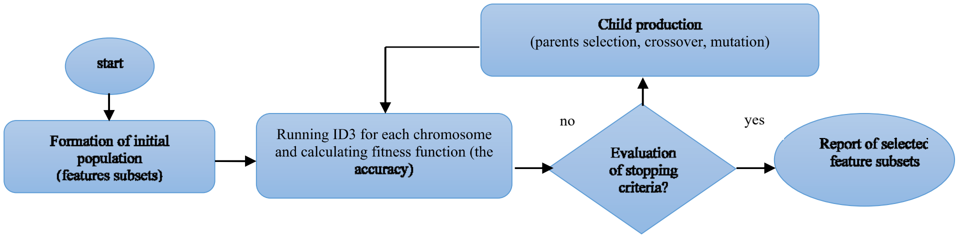
$$Entropy_A(D) = \sum_{j=1}^{m} \frac{|D_j|}{|D|} \times Entropy(D_j) \tag{2}$$

Finally, the IG value of feature A can be derived by formula (3):

$$Information\ Gain(A) = Entropy(D) - Entropy_A(D) \tag{3}$$

Each feature with more information gain is selected as the root of the tree and branched to the number of members in the domain of the feature. Then the information gain is calculated for the remaining features. This process continues until the information gain does not exceed the user's specified value (zero) or all the remaining samples belonging to the same class.

Figure 2 Feature selection using GA and ID3 (see online version for colours)


Child production
(parents selection, crossover, mutation)
start
Formation of initial population (features subsets)
Running ID3 for each chromosome and calculating fitness function (the accuracy)
no
yes
Evaluation of stopping criteria?
Report of selectec feature subsets

Source: Khan and Baig (2015)

In the next stage, if the stopping criterion is not met, the parents are selected by the tournament operator and then the two crossover and mutation operators create new subsets. These subsets are ranked again using the ID3 algorithm; this process continues until the predetermined criteria (number of generations) is met (Bhanu and Lin, 2003).

3.4 CBMA rule mining

In defining association rule mining, it is assumed that $I = \{I_1, I_2,\ldots, I_n\}$ is a set of items and n is the number of items. Also, all the samples are represented as $DB = \{T_1, T_2,\ldots, T_m\}$ in which $T_i \subseteq I$. Then the association rule mining is described as $X \rightarrow Y$ that is $X \subseteq I$, $Y \subseteq I$ and $X \cap Y = \emptyset$ (Al-Hamodi et al., 2016).

After defining the concept of association rule mining, classification based on association rule mining has been used by researchers for about two decades (Hadi et al., 2016). For the first time, Liu et al. (1998) performed the classification rule mining using the Apriori algorithm that was previously applied in association rule mining. Classification based on association rule mining like association rule mining shows the relationship between the items as if-then statements. However, in the classification, the items of the 'then' section have been predefined and they represent one class (Cano et al., 2013). In this respect, X contains one or more features and Y is a class. On the other hand, the association rule with the features from one dimension (such as purchasing) is known as a one-dimensional rule, and the rule with features from different dimensions (such as purchasing, income, age, etc.) is known as a multi-dimensional rule (Sridevi and Ramaraj, 2013). Hence, we have different features from different dimensions. On the other hand, we have specific features in each rule and particular classes. Therefore, we use the CBMA rule mining.

According to Figure 1, after feature selection, CBMA is performed and presented below (Hadi et al., 2016; Hu et al., 2016; Sridevi and Ramaraj, 2013; Liu et al., 1998):

Initial assumptions and definitions:

1 There is a database called D, in which $A = \{a_1, a_2,\ldots,a_n\}$ is the set of all features and Y is a set of all classes in D.

2 Features that belong to A are of different dimensions; therefore, all the quantitative values are divided into discrete values and represented by qualitative variables. Each feature is defined by m qualitative variable. $A' = \{(a_1, v_1), (a_1, v_2),\ldots, (a_1, v_m),\ldots, (a_n, v_m)\}$ is considered in which v_i represents a qualitative variable and $1 \leq i \leq m$.

3 A Rule item is named '*k – Ruleitem*' which is defined as '< *condset*, *y*>' where '*condset*' represents the k member set of features in A' set and $y \in Y$.

4 '*CondSupCount*' variable represents the number of samples in D database that include '*condset*'.

5 '*RuleSupCount*' variable represents the number of samples in D database that include '*condset*' with the label class of y.

6 Support is equal to $\frac{RuleSupCount}{|D|}\times 100$ where $|D|$ is the total number of samples.

7 Confidence is equal to $\frac{RuleSupCount}{CondSupCount} \times 100$.

8 The minimum support is shown by '*MinSupp*' and the minimum confidence by '*MinConf*'.

Algorithm steps:

First phase:

Step 1 All '1– *Ruleitem*' are found.

Step 2 The support is calculated for all '1– *Ruleitem*'.

Step 3 If the support of each '1 – *Ruleitem*' is equal or larger than the '*MinSupp*', it remains as '*Frequent – Ruleitem*'; otherwise, it will be removed.

Step 4 By joining '1 – *Ruleitem*' selected using the previous step, '2 – *Ruleitem*' are created.

Step 5 If after joining two '1 – *Ruleitem*'s, there are two similar dimensions in one '2 – *Ruleitem*', the '2 – *Ruleitem*' will be removed; otherwise, the support is calculated for the '2 – *Ruleitem*'.

Step 6 If the support of each '2 – *Ruleitem*' is equal or larger than the '*MinSupp*', it remains as '*Frequent – Ruleitems*'; otherwise, it will be removed.

Step 7 This search continues until the end of k^{th} level using steps (1) to (6) and after identifying all '*Frequent – Ruleitems*', phase two is performed.

Second phase:

Step 1 For all the '*Frequent – Ruleitems*' described in previous step, the confidence is calculated.

Step 2 If the confidence of each '*Frequent – Ruleitem*' is equal or larger than the '*MinConf*', *Frequent – Ruleitem* is selected as the '*Final Rule*'.

3.4.1 Rule sorting and pruning

After rule mining using CBMA, according to Figure 1, rule pruning is performed. It is done because too many rules are discovered using the CBMA algorithm and the number of rules would be reduced by pruning. For pruning, the rules are first sorted (Wedyan, 2014); then, all redundant rules are pruned according to the following steps (Thabtah, 2007):

1 Rule sorting

If there is one set of rules $R = \{R_1, R_2, \ldots, R_n\}$ from which two rules of R_a and R_b are selected, $R_a > R_b$ if:

Step 1 R_a confidence is bigger than R_b.

Step 2 The confidence of both rules is the same but the support of R_a is bigger than R_b.

Step 3 Both values of support and confidence are the same, the rule with fewer left-hand features is considered the best.

2 *Rule pruning*

Step 1 R = set of rules.

Step 2 R' = set of sorted rules.

Step 3 Remove all rules $x' \rightarrow y$ from R' where there is some rule $x \rightarrow y$ from a higher rank and $x \subseteq x'$.

3.4.2 Rule evaluation

According to Figure 1, the rules are finally evaluated after finding the final set of rules using the test dataset. One of the evaluation criteria for classification based on association rule mining is Laplace accuracy, which is calculated using formula (4). In this equation, $P_c(r)$ is the number of testing samples that correspond to r rule, $P_{tot}(r)$ is the number of testing samples that corresponds to the 'if' section of the rule, and m is the number of classes (Wedyan, 2014).

$$Laplace(r) = \frac{(P_c(r)+1)}{(P_{tot}(r)+m)} \tag{4}$$

4 Case study

Dr. Sheikh Pediatric Hospital was founded in Mashhad, Iran, in 1973 with four separate wards: nephrology, hematology, emergency, and PICU for children under 18 years old. In this study, data on the number of patients have been collected daily for 19 months from nephrology, hematology, emergency, and PICU wards. Additionally, the air pollution data were obtained from Mashhad Environmental Pollution Center, and the temperature, relative humidity, wind speed, and air pressure were all obtained from Mashhad Meteorological Ward. It should be noted that four databases for each hospital ward were considered, each of which has 581 samples and 12201 data. We extract the rules between the number of patients and different climatic factors i.e., air temperature, relative humidity, wind speed, air pressure, and air pollution in instantaneous to long-term spectrum focusing on children's patients. Primarily, we consider a pediatric hospital in Mashhad and we discover the rules for four different departments of this hospital.

5 Computational results

Here, we bring the results of different parts of the research. First, we focus on data preprocessing and then explain the feature selection and extracting rules.

5.1 Data preprocessing

First of all, the missing data were considered using IBM SPSS Statistics software, version 20. The result showed that there were no missing data. Secondly, each climatic factor is divided into four indices: instantaneous index (the value of climatic factors in patient's referral day), short-term (the average of climatic factors in one week before the patient's referral), medium-term (the average of climatic factors in two weeks before patient's referral) and long-term (the average of climatic factors in one month before referral). Then, the continuous data were divided into three discrete categories: 'low', 'middle' and 'high' as shown in Table 2.

The features and class breakpoints were determined using equal interval classification method (Dougherty et al., 1995). By this method, the range of feature and class values is divided equally. The number of categories is selected by the user (in this research the number of categories is 3). For example, the values of the number of patients in the emergency ward (N.E) are divided into three categories with equal distances. In all these categories, the distance between the initial value of the interval and the final value of the interval is equal to 18. The process of discretisation is run by Rapid Miner Studio software version 7.1.

Table 2 The breakpoints of features and classes

Features and classes	*Symbol*	*LOW*	*MIDDLE*	*HIGH*
Relative humidity in instantaneous index	H0	(3.51, 35.67)	(35.68, 67.84)	(67.85, 100)
Relative humidity in short-term index	H1	(8.56, 29.10)	(29.11, 49.65)	(49.66, 70.2)
Relative humidity in medium-term index	H2	(8.95, 27.27)	(27.28, 45.6)	(45.61, 63.93)
Relative humidity in long-term index	H3	(11.25, 25.61)	(25.62, 39.98)	(39.99, 54.35)
Air pressure in instantaneous index	P0	(797.7, 837.56)	(837.57, 877.43)	(877.44, 917.3)
Air pressure in short-term index	P1	(880.21, 890.26)	(890.27, 900.32)	(900.33, 910.38)
Air pressure in medium-term index	P2	(887.13, 894.26)	(894.27, 901.4)	(901.41, 908.54)
Air pressure in long-term index	P3	(892.31, 897.28)	(897.29, 902.26)	(902.27, 907.24)
Air temperature in instantaneous index	T0	(–7, 8.66)	(8.67, 24.33)	(24.34, 40)
Air temperature in short-term index	T1	(3.71, 15.19)	(15.2, 26.68)	(26.69, 38.17)
Air temperature in medium-term index	T2	(7.28, 17.38)	(17.39, 27.49)	(27.5, 37.6)
Air temperature in long-term index	T3	(9.03, 18.18)	(18.19, 27.36)	(27.37, 36.54)
Wind speed in instantaneous index	W0	(0, 16.8)	(16.81, 33.61)	(33.62, 50.42)

Table 2 The breakpoints of features and classes (continued)

Features and classes	*Symbol*	*LOW*	*MIDDLE*	*HIGH*
Wind speed in short-term index	W1	(5.65, 14.57)	(14.58, 23.5)	(23.51, 32.43)
Wind speed in medium-term index	W2	(6.94, 14.40)	(14.41, 21.87)	(21.88, 29.34)
Wind speed in long-term index	W3	(9.6, 15.76)	(15.77, 21.93)	(21.94, 28.1)
Air pollution in instantaneous index	PO0	(37, 89)	(89.1, 141.1)	(141.2, 193.2)
Air pollution in short-term index	PO1	(45, 79.47)	(79.48, 113.95)	(113.96, 148.43)
Air pollution in medium-term index	PO2	(46.35, 75.28)	(75.29, 104.22)	(104.23, 133.16)
Air pollution in long-term index	PO3	(53.76, 74.21)	(74.22, 94.67)	(94.68, 115.13)
The number of patients in emergency ward	N.E	(7, 25)	(26,44)	(45, 63)
The number of patients in hematology ward	N.H	(17, 26)	(27, 36)	(37, 46)
The number of patients in nephrology ward	N.N	(3, 8)	(9,14)	(15, 20)
The number of patients in PICU ward	N.P	(0, 4)	(5, 9)	(10, 14)

5.2 Feature selection

As stated earlier, climatic factors include temperature, relative humidity, wind speed, air pressure, and air pollution. Each climatic factor is divided into four indices: instantaneous, short-, medium- and long-term. In this way, 5 features have increased to 20 features. After grouping the features and discretising the numerical values, feature selection is applied for the number of patients in each hospital ward. The reason that feature selection was done after grouping climatic factors into 4 indices is that according to the literature review and experts' opinion, all climatic factors affect the number of patients, but their effectiveness may be various in different periods.

Table 3 The best obtained GA parameters in different hospital wards

	P	*G*	*T*	*C*	*M*	*A*
Emergency ward	20	400	0.5	0.7	0.1	*75.29*
Hematology ward	20	30	0.5	0.5	0.1	*86.78*
Nephrology ward	20	100	0.5	0.5	0.05	*61.49*
PICU ward	15	400	0.6	0.6	0.05	*77.01*

Table 4 Feature selection in each hospital ward

The number of patients in deferent wards	*Climatic factors*																			
	H0	*H1*	*H2*	*H3*	*P0*	*P1*	*P3*	*P2*	*T0*	*T1*	*T2*	*T3*	*W0*	*W1*	*W2*	*W3*	*PO0*	*PO1*	*PO2*	*PO3*
N.E	√	√		√		√		√				√				√		√		
N.H		√	√	√	√	√		√	√			√				√		√	√	√
N.N		√		√	√	√	√	√	√	√		√		√	√	√	√	√	√	√
N.P	√	√	√	√		√	√	√	√	√	√	√	√	√	√	√	√			√

In this research, feature selection was performed by Rapid Miner Studio software version 7.1. In this software, GA was used for generating the subsets and the ID3 decision trees were selected for evaluating each subset. In other words, the algorithm described in the 3.3.1 section has been run by Rapid Miner Studio software. The software automatically generates decision trees by randomly selecting 70% of the data and it measures the accuracy of each tree by considering30% of the data. The accuracy calculation in this software is based on cross-validation, in which the data is divided into k subsets, in which every subset is used for validation and k-1 subset for training. This process is repeated K times and all data are used once for the training and once for validation. Finally, the average of this K-fold validation is selected as a final estimate. Feature selection was calculated separately for the number of patients in each hospital ward, and the GA in different types is performed with the population size (P) 10, 15, 20, generation number (G) 30, 50, 100, 200, 300, 400, 500, tournament size (T) 0.4, 0.5, 0.6, crossover rate (C) 0.5, 0.6, 0.7 and mutation rate (M) 0.05 and 0.1. Table 3 shows the best set of GA parameters to achieve the highest measured accuracy (A) for the number of patients in each hospital wards and Table 4 illustrates feature selection. Finally, 9 features were selected for the number of patients in emergency (N.E), 12 features for the number of patients in hematology (N.H), 16 features for the number of patients in nephrology (N.N), and 17 features for the number of patients in PICU (N.P).

5.3 CBMA rule mining

In this study, all databases were transformed into two sections: training data containing 70% of the data and testing databases containing 30% of data using IBM SPSS Modeler software version 18. The training databases were used to implement the CBMA algorithm. CBMA algorithm was implemented for each hospital ward by considering MinSupp = 1% and MinConf = 80%. The reason for considering low value for Support is to find rules that are less frequent and sometimes rare with high confidence. The number of rules found for N.E is 1,233, for N.H is 7,028, for N.N is 2,536 and for N.P is 16,044.

MATLAB Programming was used to sort and prune the rules in this study. After pruning the rules, the number of rules greatly decreased, and the number of rules equaled 118, 218, 709, and 1187 for N.E, N.H, N.N, and N.P respectively.

Ultimately, the final rules with Support (S), Confidence (C), and Laplace accuracy (L) are shown in Tables 5 to 8 after pruning and evaluation.

Table 5 Final rules for N.E

RULES		*S (%)*	*C (%)*	*L (%)*
1)	If P3 = LOW and H3 = LOW and W3 = MIDDLE then N.E = LOW	8.88	100	72.72
2)	If P3 = LOW and H3 = LOW and PO1 = LOW then N.E = LOW	6.91	100	55.55
3)	If PO3 = HIGH and H0 = MIDDLE then N.E = MIDDLE	3.21	100	71.42
4)	If T3 = MIDDLE and P1 = MIDDLE and H1 = LOW and PO3 = MIDDLE then N.E = HIGH	2.21	100	50
…		…	…	
117)	If T3 = MIDDLE and P1 = MIDDLE and PO1 = LOW and PO3 = MIDDLE then N.E = HIGH	2.46	80	50
118)	If PO1 = HIGH and H0 = MIDDLE then N.E = MIDDLE	1.23	80	50

Table 6 Final rules for N.H

RULES		*S (%)*	*C (%)*	*L (%)*
1)	If H3 = MIDDLE and W3 = LOW and PO3 = MIDDLE then N.H = MIDDLE	15.80	100	93.54
2)	If W3 = LOW and H2 = MIDDLE and PO3 = MIDDLE then N.H = MIDDLE	15.06	100	88.88
3)	If W3 = HIGH and P1 = HIGH then N.H = HIGH	2.96	100	75
4)	If T3 = MIDDLE and H3 = LOW and PO1 = LOW then N.H = HIGH	2.96	100	71.42
…		…	…	
217)	If T3 = MIDDLE and P3 = MIDDLE and W3 = MIDDLE and PO2 = MIDDLE then N.H =LOW	1.23	80	40
218)	If H3 = HIGH and H2 = MIDDLE and PO1 = MIDDLE and P3 = MIDDLE then N.H = LOW	1.23	80	40

Table 7 Final rules for N.N

RULES		*S (%)*	*C (%)*	*L (%)*
1)	If H3 = HIGH and PO1 = MIDDLE and P2 = HIGH and W2 = MIDDLE then N.N =HIGH	2.46	100	62.5
2)	If H3 = HIGH and PO1 = MIDDLE and W2 = MIDDLE and P1 = HIGH then N.N =HIGH	2.46	100	62.5
3)	If PO3 = LOW and W3 = LOW and P3 = HIGH then N.N = LOW	1.48	100	60
4)	If H3 = HIGH and T1 = MIDDLE and H1 = MIDDLE and W3 = LOW N.N = LOW	1.48	100	42.85
…		…	…	
708)	If PO2 = LOW and P3 = MIDDLE and P1 = HIGH and PO3 = MIDDLE then N.N =MIDDLE	1.23	80	50
709)	If W2 = LOW and H3 = LOW and T0 = HIGH and PO3 = MIDDLE then N.N = MIDDLE	1.23	80	40

Table 8 Final rules for N.P

RULES		*S (%)*	*C (%)*	*L (%)*
1)	If T3 = LOW and P3 = HIGH and W1 = MIDDLE then N.P = HIGH	6.66	100	83.33
2)	If W3 = HIGH and PO3 = LOW then N.P = HIGH	6.42	100	90
3)	If P3 = LOW and H2 = MIDDLE then N.P = LOW	1.48	100	60
4)	If H2 = MIDDLE and T3 = HIGH and W3 = MIDDLE and PO3 = MIDDLE then N.P = LOW	1.23	100	60
…		…	…	
1186)	If T1 = MIDDLE and W1 = LOW and H3 = LOW and T0 = HIGH then N.P = MIDDLE	1.23	80	60
1187)	If H0 = MIDDLE and H1 = MIDDLE and W0 = MIDDLE and W3 = MIDDLE then N.P = MIDDLE	1.23	80	42.85

6 Discussion

In this study, the relationship between climatic factors such as relative humidity, air pressure, wind speed, temperature, and air pollution in four instantaneous, short-, medium- and long-term indices was evaluated with the number of hospital patients. In the first phase of research, climatic factors were selected for each of these indices using the GA and ID3 decision tree. The results showed that all climatic factors were correlated with the number of patients in different hospital wards. Especially, there was a relationship between climatic factors in the long-term index with the number of patients in different wards. However, the relationship between climatic factors in other indices and the number of patients was different.

The effect of climatic factors on the number of patients in different wards in this research is in line with the results of Tauler et al. (1985), D'Souza et al. (2008), Mireku et al. (2009), Onozuka and Hashizume (2011), Kim et al. (2017), Martín Martín and Sánchez Bayle (2018), Moradiasl et al. (2018), Yu et al. (2018), Liu et al. (2019) and Li et al. (2020). In all of the previous studies, the effect of one or more climatic factors on diseases has been proven. However, in previous studies, only positive and negative correlations between diseases and climatic factors have been studied. In this research, all climatic factors have been studied, and the climatic factors have also been transformed into four indices: instantaneous, short-, medium- and long-term. This transformation has made it possible to analyse the impact of each climatic factor on the number of patients at different times. A climatic factor may have different effects on diseases and the number of patients at different times, which has not been considered in previous studies.

The relationship between climatic factors in each hospital ward is shown in Figures 3 to 6. In these figures, 50 strong links of climatic factors are selected; the bold lines indicate stronger connections. Using these figures, we can see which features are along with each other and more often repeated. For example, in Figure 3 H3 = low has a strong connection with H1 = low, H0 = low, and T3 = high.

Figure 3 Strong links between features in N.E (see online version for colours)

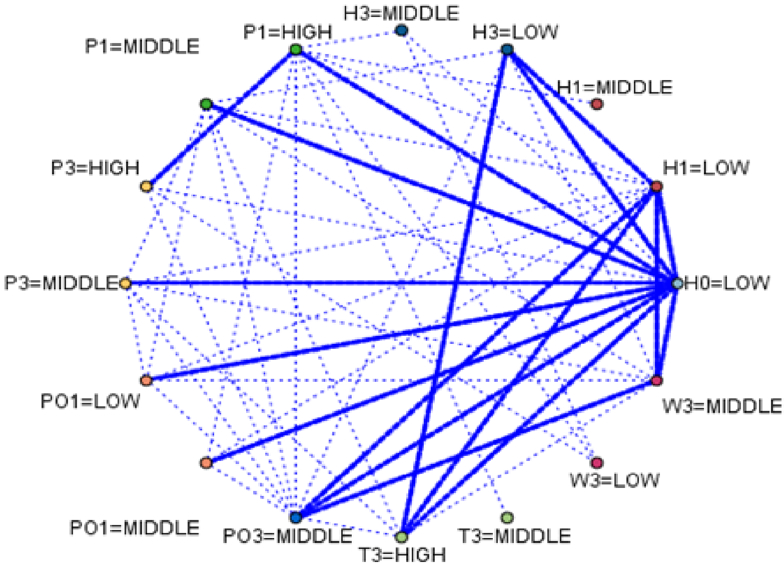


Figure 4 Strong links between features in N.H (see online version for colours)

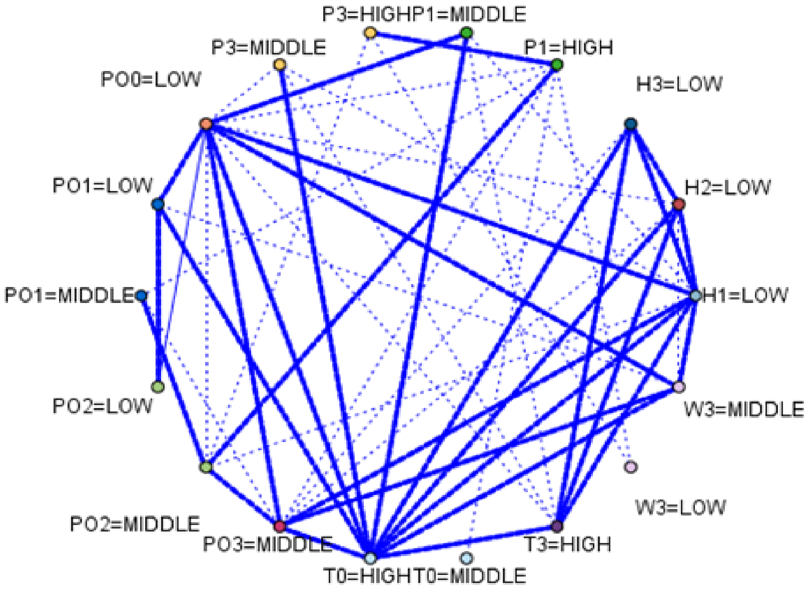


Figure 5 Strong links between features in N.N (see online version for colours)

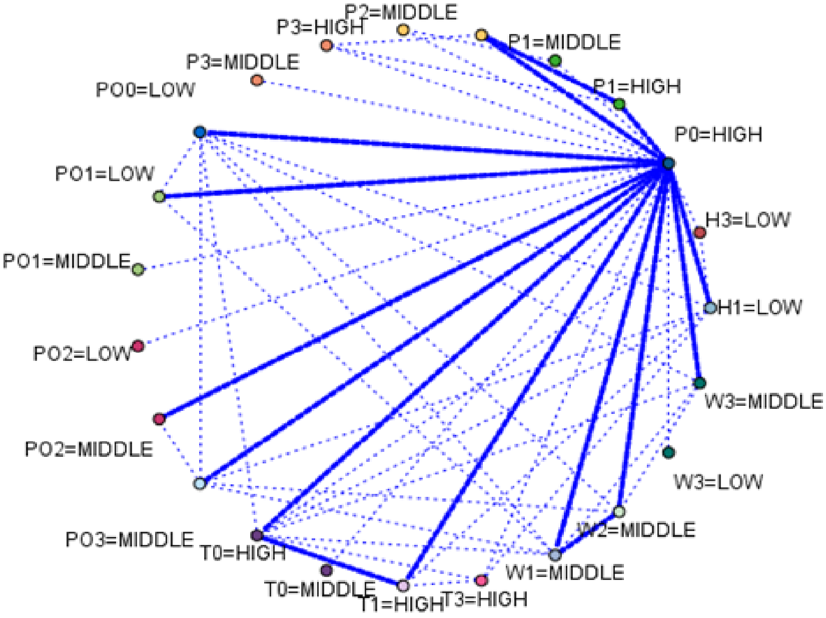


Figure 6 Strong links between features in N.P (see online version for colours)

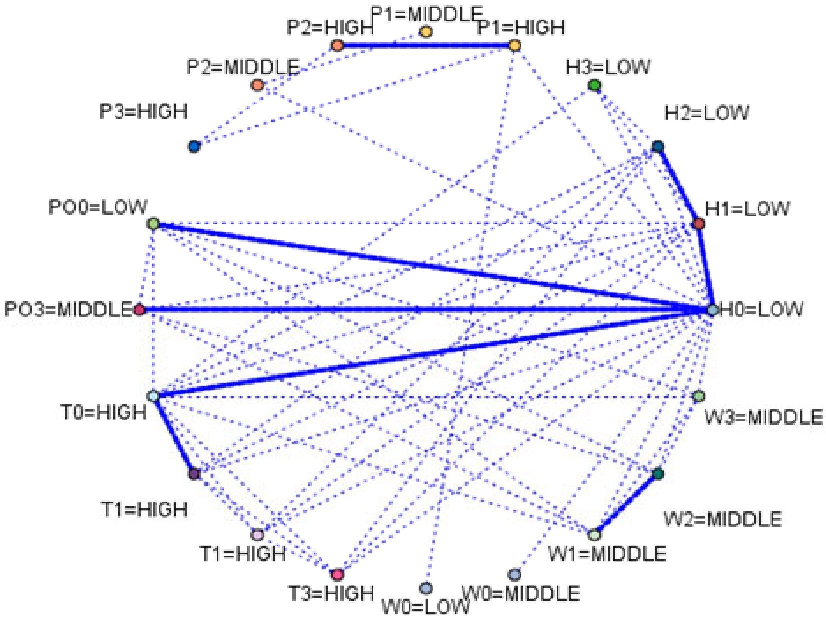
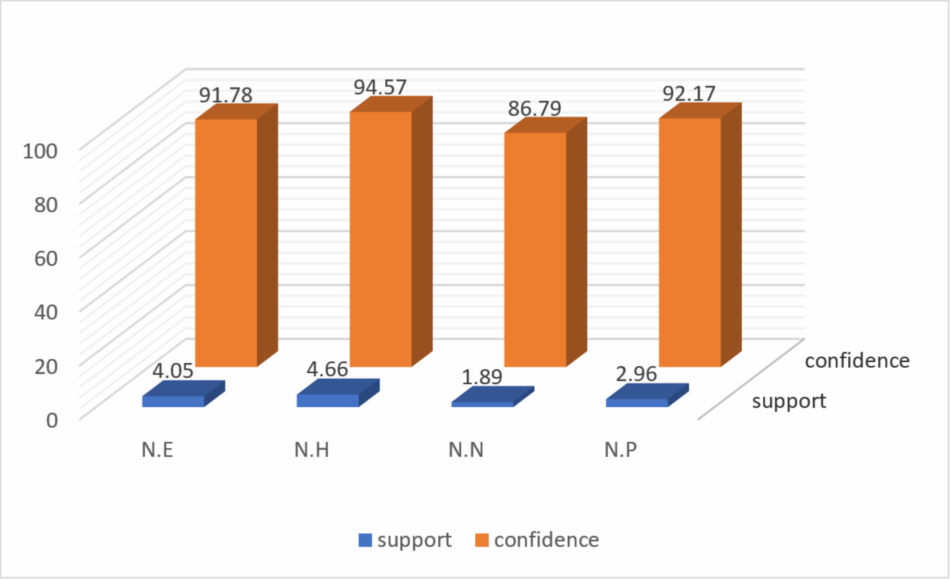


Figure 7 The support and confidence average of rules in different wards (see online version for colours)



In the second phase, the number of patients in each hospital ward was determined using CBMA, and the accuracy of the rules was evaluated using Laplace accuracy. Because of using different climatic factors in four indices, the rules obtained in this study had a more comprehensive result in comparison with the studies carried out by Martín Martín and Sánchez Bayle (2018), and Toti et al. (2016), who have predicted the changes in the number of patients (children) as a result of only air pollution factor. Since the combination of methods and datasets used in this research has not been used in previous studies, it is not possible to make quantitative comparisons between the results of this research and related works.

In this study, different rules for the number of patients in each hospital ward with varying levels of support and confidence were obtained. In Figure 7, the average of support and confidence for the rules of four hospital wards are compared.

As shown in the above figures, the rules predicting the number of patients in the hematology ward has the highest average of support and confidence, and the rules predicting the number of patients in nephrology has the lowest average of support and confidence. Also, the support and confidence maximum and minimum are shown in Figures 8 and 9. For example, in Figures 7 to 9, the support and confidence average of rules in N.E is 4.05% and 91.78%; The support and confidence maximum of rules in N.E is 21.48 % and 100%; The support and confidence minimum of rules in N.E is 1.23% and 80%.

Figure 8 The support and confidence maximum of rules in different wards (see online version for colours)

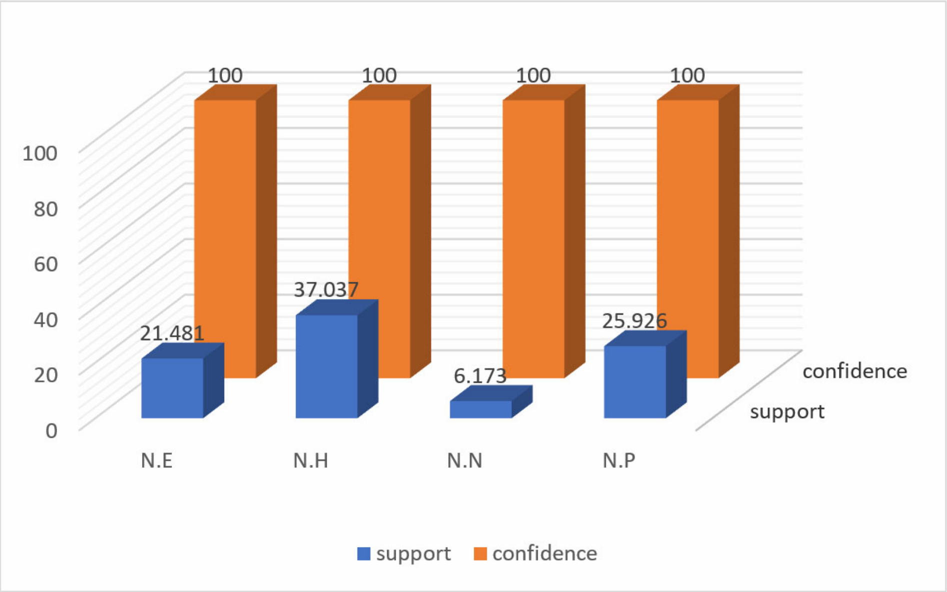
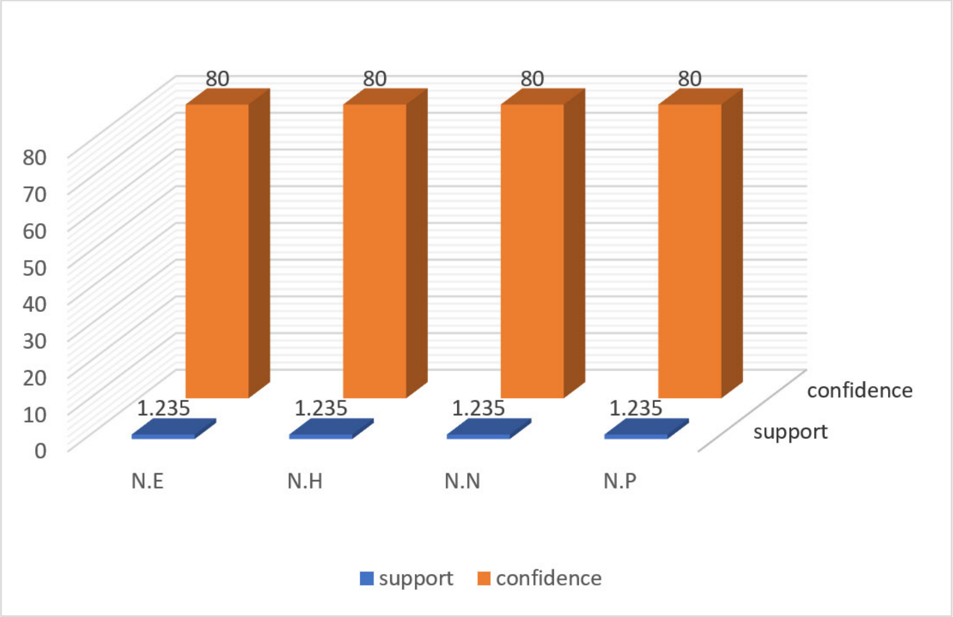


Figure 9 The support and confidence minimum of rules in different wards (see online version for colours)



As stated in the previous sections of this research, the obtained rules can be used to predict the number of patients in the coming days. The rules extracted are a quantitative way to help physicians, nurses, and decision-makers make better decisions and provide quick and inexpensive estimates of the number of patients to plan for the shortage or excess of hospital resources.

7 Conclusions

In this study, the relationship between climatic factors in four indices of instantaneous, short-, medium- and long-term with the number of patients is generally obtained by data mining and in particular by GA, ID3, and CBMA. The results showed that in the long-term index, all climatic factors were correlated with the number of patients in all hospital wards, but in the other three indices (instantaneous, short- and medium-term), some climatic factors were involved. More importantly, the results suggest many rules that illustrate the changes in the number of patients as low, medium, and high due to climatic factors in different hospital wards.

One of the main applications of this study is to predict the number of patients on different days according to climatic factors using the founded rules. Predicting the number of patients is very helpful in eliminating the mismatch between the demand and supply of human and financial resources of the hospital. This leads to the optimal allocation of hospital resources, more accurate planning for attendance time of doctors and nurses, reduction of unnecessary costs and reduction of waiting time for patients. This research showed that since it is possible to estimate the number of children referred to different parts of the hospital on different days depending on climatic factors, it is crucial to consider these rules for proper resource allocation in hospitals.

The limitation of the present study has been the lack of data availability. There were not enough data in the case study for many years, which can affect the accuracy of the results. It is suggested that future research should focus on comparing the results of this algorithm with other classification algorithms such as C5.0 decision tree classification algorithm, CART decision tree, hybrid classifications, and random forest algorithm, etc.

References

Alaiad, A., Najadat, H., Mohsen, B. and Balhaf, K. (2020) 'Classification and association rule mining technique for predicting chronic kidney disease', *Journal of Information & Knowledge Management*, Vol. 19, No. 1, p.2040015.

Al-Hamodi, A.A.G., Lu, S. and Al-Salhi, Y.E.A. (2016) 'An enhanced frequent pattern growth based on MapReduce for mining association rules', *International Journal of Data Mining & Knowledge Management Process*, Vol. 6, No. 2, pp.19–28.

Alwidian, J., Hammo, B.H. and Obeid, N. (2018) 'WCBA: weighted classification based on association rules algorithm for breast cancer disease', *Applied Soft Computing*, Vol. 62, pp.536–549.

Bhanu, B. and Lin, Y. (2003) 'Genetic algorithm-based feature selection for target detection in SAR images', *Image and Vision Computing*, Vol. 21, No. 7, pp.591–608.

Borah, A. and Nath, B. (2018) 'Identifying risk factors for adverse diseases using dynamic rare association rule mining', *Expert Systems with Applications*, Vol. 113, pp.233–263.

Cano, A., Zafra, A. and Ventura, S. (2013) 'An interpretable classification rule mining algorithm', *Inf. Sci.*, Vol. 240, pp.1–20.

Chandrashekar, G. and Sahin, F. (2014) 'A survey on feature selection methods', *Computers & Electrical Engineering*, Vol. 40, No. 1, pp.16–28.

D'Souza, R.M., Hall, G. and Becker, N.G. (2008) 'Climatic factors associated with hospitalizations for rotavirus diarrhoea in children under 5 years of age', *Epidemiol Infect*, Vol. 136, No. 1, pp.56–64.

Dash, M. and Liu, H. (2003) 'Consistency-based search in feature selection', *Artificial Intelligence*, Vol. 151, Nos. 1–2, pp.155–176.

Dougherty, J., Kohavi, R. and Sahami, M. (1995) 'Supervised and unsupervised discretization of continuous features', *Machine Learning Proceedings 1995*, Elsevier.

Hadi, W.E., Aburub, F. and Alhawari, S. (2016) 'A new fast associative classification algorithm for detecting phishing websites', *Applied Soft Computing*, Vol. 48, pp.729–734.

Hervás, D., Utrera, J.F., Hervás-Masip, J., Hervás, J.A. and García-Marcos, L. (2015) 'Can meteorological factors forecast asthma exacerbation in a paediatric population?', *Allergologia et Immunopathologia*, Vol. 43, No. 1, pp.32–36.

Hu, L-Y., Hu, Y-H., Tsai, C-F., Wang, J.-S. and Huang, M-W. (2016) 'Building an associative classifier with multiple minimum supports', *SpringerPlus*, Vol. 5, pp.528–528.

Ivancevic, V., Tusek, I., Tusek, J., Knezevic, M., Elheshk, S. and Lukovic, I. (2015) 'Using association rule mining to identify risk factors for early childhood caries', *Comput. Methods Programs Biomed.*, Vol. 122, No. 2, pp.175–81.

Javed, K., Babri, H.A. and Saeed, M. (2012) 'Feature selection based on class-dependent densities for high-dimensional binary data', *IEEE Trans. on Knowl. and Data Eng.*, Vol. 24, No. 3, pp.465–477.

Khan, A. and Baig, A.R. (2015) 'Multi-objective feature subset selection using non-dominated sorting genetic algorithm', *Journal of Applied Research and Technology*, Vol. 13, No. 1, pp.145–159.

Kim, S.H., Kim, J.S., Jin, M.H. and Lee, J.H. (2017) 'The effects of weather on pediatric seizure: a single-center retrospective study (2005-2015)', *Sci. Total Environ.*, Vol. 609, pp.535–540.

Lai, C-M., Yeh, W-C. and Chang, C-Y. (2016) 'Gene selection using information gain and improved simplified swarm optimization', *Neurocomputing*, Vol. 218, pp.331–338.

Li, Y., Ye, X., Zhou, J., Zhai, F. and Chen, J. (2020) 'The association between the seasonality of pediatric pandemic influenza virus outbreak and ambient meteorological factors in Shanghai', *Environmental Health*, Vo.19, No. 1, pp.1–10.

Liu, B., Hsu, W. and Ma, Y. (1998) *Integrating Classification and Association Rule Mining*, pp.80–86, KDD.

Liu, W.K., Chen, D.H., Tan, W.P., Qiu, S.Y., Xu, D., Zhang, L., Gu, S.J., Zhou, R. and Liu, Q. (2019) 'Paramyxoviruses respiratory syncytial virus, parainfluenza virus, and human metapneumovirus infection in pediatric hospitalized patients and climate correlation in a subtropical region of southern China: a 7-year survey', *European Journal of Clinical Microbiology & Infectious Diseases*, Vol. 38, No. 12, pp.2355–2364.

Martín Martín, R. and Sánchez Bayle, M. (2018) 'Impact of air pollution in paediatric consultations in primary health care: ecological study', *Anales de Pediatría (English Edition)*, Vol. 89, No. 4, pp.80–85.

Mireku, N., Wang, Y., Ager, J., Reddy, R.C. and Baptist, A.P. (2009) 'Changes in weather and the effects on pediatric asthma exacerbations', *Ann Allergy Asthma Immunol.*, Vol. 103, No. 3, pp.220-224.

Moradiasl, E., Rassi, Y., Hanafi-Bojd, A.A., Vatandoost, H., Saghafipour, A., Adham, D., Aabasgolizadeh, N., Omidi Oskouei, A. and Sadeghi, H. (2018) 'The relationship between climatic factors and the prevalence of visceral leishmaniasis in North West of Iran', *International Journal of Pediatrics*, Vol. 6, No. 2, pp.7169–7178.

Nahar, J., Imam, T., Tickle, K.S. and Chen, Y-P.P. (2013) 'Association rule mining to detect factors which contribute to heart disease in males and females', *Expert Systems with Applications*, Vol. 40, No. 4, pp.1086–1093.

North, M. (2012) *Data Mining for the Masses*, Lexington, Ky., Global Text Project, USA.

Onozuka, D. and Hashizume, M. (2011) 'The influence of temperature and humidity on the incidence of hand, foot, and mouth disease in Japan', *Sci. Total Environ.*, Vols. 410–411, pp.119–25.

Ordonez, C. (2006) 'Association rule discovery with the train and test approach for heart disease prediction', *IEEE Transactions on Information Technology in Biomedicine*, Vol. 10, pp.334–343.

Ordonez, C., Omiecinski, E., Braal, L.D., Santana, C.A., Ezquerra, N., Taboada, J.A., Cooke, D., Krawczynska, E. and Garcia, E.V. (2001) 'Mining constrained association rules to predict heart disease', *Proceedings 2001 IEEE International Conference on Data Mining*, 29 November–2 December, pp.433–440.

Sarıyer, G. and Öcal Taşar, C. (2020) 'Highlighting the rules between diagnosis types and laboratory diagnostic tests for patients of an emergency department: use of association rule mining', *Health Informatics Journal*, Vol. 26, No. 2, pp.1177–1193.

Sridevi, R. and Ramaraj, E. (2013) 'A general survey on multidimensional and quantitative association rule mining algorithms', *International Journal of Engineering Research and Applications*, Vol. 3, No. 4, pp.1442–1448.

Tauler, E., Llorens-Terol, J., Mur, A. and Leal, C. (1985) 'Asthma and environmental factors', *Pediatric Research*, Vol. 19, No. 10, pp.1120–1120.

Thabtah, F. (2007) 'A review of associative classification mining', *The Knowledge Engineering Review*, Vol. 22, No. 1, pp.37–65.

Toti, G., Vilalta, R., Lindner, P., Lefer, B., Macias, C. and Price, D. (2016) 'Analysis of correlation between pediatric asthma exacerbation and exposure to pollutant mixtures with association rule mining', *Artificial Intelligence in Medicine*, Vol. 74, pp.44–52.

Wang, C.H., Lee, T.Y., Hui, K.C. and Chung, M.H. (2019) 'Mental disorders and medical comorbidities: association rule mining approach', *Perspectives in Psychiatric Care*, Vol. 55, No. 3, pp.517–526.

Wedyan, S. (2014) 'Review and comparison of associative classification data mining approaches', *International Journal of Computer, Information, Systems and Control Engineering*, Vol. 8, No. 1, pp.34–45.

Weng, C-H. (2016) 'Identifying association rules of specific later-marketed products', *Applied Soft Computing*, Vol. 38, pp.518–529.

Xiaohu, W., Lele, W. and Nianfeng, L. (2012) 'An application of decision tree based on ID3', *Physics Procedia*, Vol. 25, pp.1017–1021.

Yu, G., Fu, Y., Dong, C., Duan, H. and Li, H. (2018) 'Is the occurrence of pediatric epistaxis related to climatic variables?', *International Journal of Pediatric Otorhinolaryngology*, Vol. 113, pp.182–187.

Zhang, C. and Zhang, S. (2002) *Association Rule Mining: Models and Algorithms*, Springer-Verlag, Berlin, Heidelberg.