Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 9, No. 1, 2024

Air Quality Correlation Analysis Based on Apriori Algorithm

Shaohui Yan[1], Shuai Kang[1, *]

[1]School of Physics and Electronic Engineering, Northwest Normal University, Lanzhou, 10736, China
[*]Corresponding author: kangshun2024@126.com

Abstract: In recent years, air pollution has become an environmental issue of global concern. It not only affects human health, but also causes serious damage to the Earth's ecosystem. By analyzing the concentration values of various pollutants in the air and the air quality index (AQI) every hour, the most relevant components to the AQI can be identified. In this paper, Apriori association rule algorithm is applied to correlate various data in air quality monitoring to provide a new reference solution for solving environmental pollution problems in the future.

Keywords: Apriori algorithm, Association rules, Data preprocessing.

1. Introduction

Air pollution is a serious environmental problem that poses a great threat to the health of human beings and the planet. Its main sources are industrial, automobile and household emissions. These pollutants form smog, suspended particles and harmful gases in the atmosphere, posing a serious threat to human health. In order to improve air quality and effectively suppress the concentration of pollutants in the air, Apriori algorithm is proposed to analyze the correlation of air quality detection data.

At present, the research on correlation at home and abroad has achieved certain results. Literature [1] proposed a multi-objective genetic algorithm based on rough patterns for mining the association rules between values. The rough values defined by upper and lower bounds are used to represent a set of values, and Pareto optimality is used to solve the multi-objective optimization problem, which contributes well to improve the performance of the algorithm. Literature [2] selects the most effective attributes by using association rule-optimized wheeze measurements and dissolved gas analysis methods, which reduces the inputs to the adaptive neuro-fuzzy inference system and effectively improves the performance of the algorithm. Literature [3] proposed an Apriori algorithm based on association rule mining to discover symptom patterns in COVID-19 patients. This study used records of 2875 patients to identify the most common signs and symptoms, providing clinicians with valuable insights into the disease and helping them to effectively manage and treat it.

In this paper, an Apriori algorithm based correlation study of air quality testing data is proposed. First, the obtained data are processed to obtain a set of basic ground correlation data set. Then, the data set is analyzed using Apriori algorithm to get the correlation results between the data. Finally, the accuracy of the correlation results is examined. In this paper, a correlation study of air quality detection data based on Apriori algorithm is proposed. First, the obtained data are processed to obtain a set of basic ground correlation data set. Then, the data set is analyzed using Apriori algorithm to get the correlation results between the data. Finally, the accuracy of the correlation results is examined.

2. The Apriori Algorithm

2.1. Basic concepts

Association rules are a very classical task in data mining, whose main goal is to extract frequent items and corresponding association rules from a series of transaction sets and to explore the interdependence and correlation between events and other events. The association rule mining algorithms proposed so far are Apriori algorithm, FP growth algorithm, Eclat algorithm, etc [5-7]. One of the most famous association rule methods is Apriori algorithm .

Before gaining a specific understanding of the Apriori algorithm, it is important to first understand some concepts:

①Association rule: An association rule is an implication in the form of X=>Y, where X and Y are itemsets, and there is no intersection between X and Y. X is called the antecedent of the rule, and Y is called the consequent of the rule. The association rule reflects the pattern in which the items in X appear, and the items in Y also follow.

②Support: The proportion of times several associated data appear in the dataset to the total dataset [10]. Namely, support=(including the number of records for item X)/(total number of records) [9]

$$\text{support}\ (x, y) = P(XY) = \frac{number(XY)}{number(AllSamples)} \quad (1)$$

③Confidence: The probability of one data appearing, or the conditional probability of another data appearing. Confidence (X ->Y)=(number of records containing items X and Y)/(number of records containing X). Confidence is a conditional concept, which means what is the probability of Y occurring in the case of X [11].

$$confidence(X->Y) = P(Y \mid X) = \frac{P(XY)}{P(X)} \quad (2)$$

④Minimum support and minimum confidence: Usually, users need to specify the support and confidence thresholds that the rule must meet in order to meet certain requirements.

When support (X=>Y) and confidence (X=>Y) are greater than or equal to their respective threshold values, X=>Y is considered meaningful, and these two values are called the minimum support threshold (min_sup) and minimum confidence threshold (min_conf). Among them, min_ sup describes the minimum importance level of association rules, min_conf specifies the minimum reliability that association rules must meet.

2.2. Main steps

The Apriori algorithm [8] is an iterative method of layer by layer search used to explore the correlations between datasets. It first needs to find the set L1 of frequent 1-itemsets, then use L1 to find the set L2 of frequent 2-itemsets, and so on, until the frequent K-itemsets cannot be found. There are two main steps in this, which are connecting and pruning.

①Connection: Using the frequent itemsets of k items that have already been found, a new candidate itemset is obtained by pairwise combination connection. Note that the two different frequent itemsets connected at this time must have k-1 items that are the same.

②Pruning: Not all candidate itemsets obtained are frequent itemsets, and pruning operations must be performed to obtain truly frequent itemsets with candidate support greater than the minimum support.

2.3. Related Examples

The itemset processed by the Apriori algorithm, if a certain itemset is frequent, then all its subsets are also frequent. Specific examples are shown in Figure 2

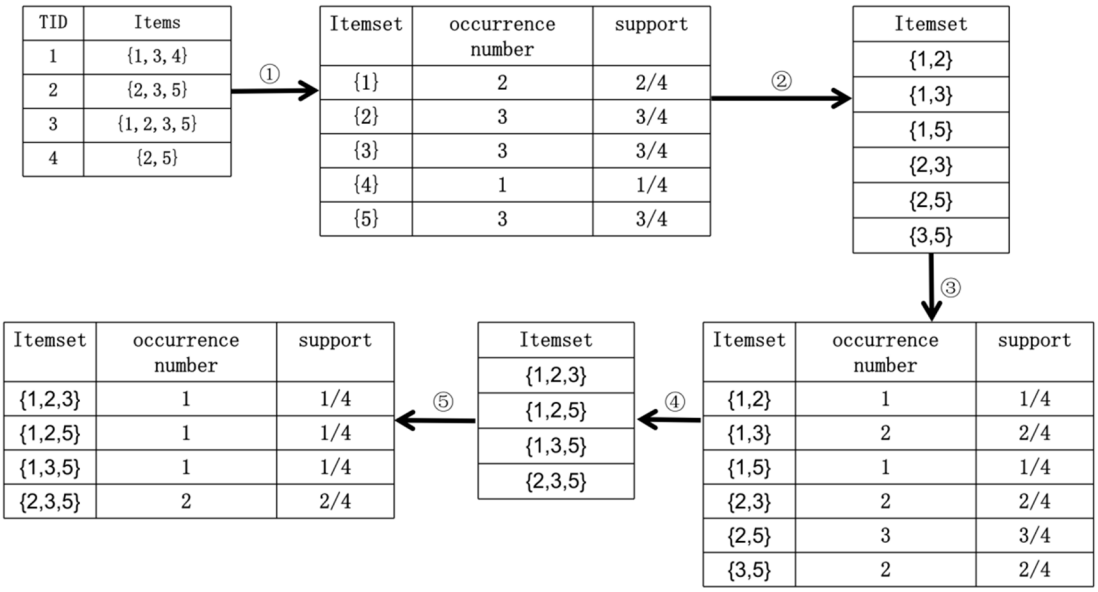

Figure 1. Related Examples

There are a total of 4 records in the dataset, namely {1,3,4}, {2,3,5}, {1,2,3,5}, and {2,5}. The minimum support threshold (min_sup) is set to 50%

① Generate candidate frequent itemsets for the dataset and calculate their respective support levels (4 records, where 1, 2, 3, 4, and 5 appear as 2, 3, 3, 1, and 3 respectively)

② Cut out candidate frequent items with support less than 50%, and combine the remaining frequent itemsets pairwise to generate candidate frequent 2-itemsets (combine the remaining itemsets by pruning according to step ①). At this point, the first round of iteration ends

③ Enter the second iteration and calculate the support levels for the generated frequent 2-item sets

④Cut out candidate frequent items with support less than 50%, and combine the remaining frequent itemsets in pairs to generate candidate frequent itemsets of 3. At this point, the second round of iteration ends

⑤ Entering the third round of iteration, for the candidate frequent 3-itemset, calculate the support for each item, trim out itemsets less than 50%, and finally leave {2,3,5}. At this point, the number of frequent items is 1, and generating candidate frequent 4-itemsets is not supported. The iteration ends.

3. Empirical Results

3.1. Data acquisition

Air Quality Index (AQI) [4] is a parameter that reflects the daily air quality. It is calculated based on the measured concentration values of various pollutants (e.g., fine particulate matter (PM2.5), respirable particulate matter (PM10), sulfur dioxide (SO2), nitrogen dioxide (NO2), ozone (O3), carbon monoxide (CO), etc.). The air quality measurement data are available on the Air Quality Monitoring and Analysis Platform (AQMA) for each time period.

3.2. Data processing

Due to the presence of some empty values in the obtained data, it is necessary to process the data. The main processing methods include: supplementing with the last value, supplementing with the first value, supplementing with the average value, and removing all data in the row where the empty values are located. This article chooses the processing method of removing all data in the row where the null value is located. The processing effect is shown in Figure 2 and Figure 3.

03_24h	03_8h	03_8h_24h	CO	C
42	10	37	0. 7	
42	nan	nan	0. 7	
42	nan	nan	0. 7	
42	nan	nan	0. 7	
42	nan	nan	0. 7	
42	nan	nan	0. 7	
42	nan	nan	0. 5	
42	nan	nan	0. 7	
42	nan	nan	0. 6	
42	1	1	0. 5	
42	4	4	0. 5	

Figure 2. Pre processing data

03_24h	03_8h	03_8h_24h	CO	
42	10	37	0. 7	
42	1	1	0. 5	
42	4	4	0. 5	
42	7	7	0. 5	
42	11	11	0. 5	
35	16	16	0. 5	

Figure 3. Processed data

In addition, in order to provide the association set for the Apriori algorithm, further processing of the data is required. The processing method is to synthesize the data that continuously changes the variables within the same time period, and then generate the data set required by the Apriori algorithm.

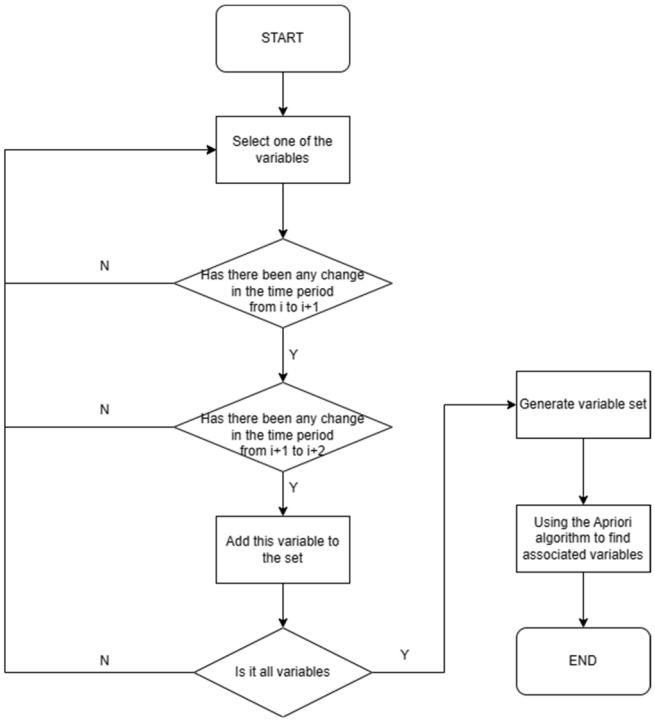

Figure 4. Algorithm flowchart

3.3. Apriori algorithm

Use the Apriori algorithm to find the relevant variables in the generated correlation set, set the minimum support to 0.01, and first perform correlation analysis on the data variable set of one time period

```
['AQI', 'PM2.5', 'PM10', 'NO2', 'O3'] 0.19683481701285854
['AQI', 'PM2.5', 'NO2', 'O3', 'CO'] 0.05016249823371485
['PM2.5', 'PM10', 'NO2', 'O3', 'CO'] 0.051575526352974424
['AQI', 'PM10', 'NO2', 'O3', 'CO'] 0.053129857284159955
['AQI', 'PM2.5', 'PM10', 'O3', 'CO'] 0.0579341528896425
['AQI', 'PM2.5', 'PM10', 'NO2', 'CO'] 0.057368941641938676
['AQI', 'PM2.5', 'PM10', 'SO2', 'O3'] 0.024727992087042534
['PM2.5', 'PM10', 'SO2', 'NO2', 'O3'] 0.016108520559559136
['AQI', 'PM2.5', 'PM10', 'SO2', 'NO2'] 0.018934576798078282
['AQI', 'PM2.5', 'SO2', 'NO2', 'O3'] 0.01695633743111488
['AQI', 'PM10', 'SO2', 'NO2', 'O3'] 0.018793273986152326
['AQI', 'PM2.5', 'SO2', 'O3', 'CO'] 0.010739013706372758
['AQI', 'PM2.5', 'PM10', 'SO2', 'CO'] 0.013423767132965945
['PM2.5', 'PM10', 'SO2', 'O3', 'CO'] 0.010456408082520843
['AQI', 'PM10', 'SO2', 'O3', 'CO'] 0.011586830577928501
['AQI', 'PM10', 'SO2', 'NO2', 'CO'] 0.010173802458668928
['AQI', 'PM2.5', 'PM10', 'NO2', 'O3', 'CO'] 0.04804295605482549
['AQI', 'PM2.5', 'PM10', 'SO2', 'NO2', 'O3'] 0.01582591493570722
['AQI', 'PM2.5', 'PM10', 'SO2', 'O3', 'CO'] 0.010456408082520843
```

Figure 5. Correlation Results for Partial Time Periods

As shown in Figure 5, the relationship between AQI and PM2.5, PM10, NO2, O3 is relatively close. Next, the minimum support level is still set to 0.01, and data from other time periods are analyzed. As shown in Figure 6, the relationship between AQI and PM2.5, PM10, NO2, O3 is still close.

```
['AQI', 'PM2.5', 'SO2', 'NO2', 'O3'] 0.017365097588978185
['AQI', 'PM2.5', 'PM10', 'SO2', 'O3'] 0.022962112514351322
['AQI', 'PM2.5', 'PM10', 'NO2', 'O3'] 0.16776693455797934
['AQI', 'PM2.5', 'PM10', 'SO2', 'NO2'] 0.019230769230769232
['PM2.5', 'PM10', 'SO2', 'NO2', 'O3'] 0.016791044776119403
['AQI', 'PM10', 'SO2', 'NO2', 'O3'] 0.019087256027554535
['AQI', 'PM10', 'NO2', 'O3', 'CO'] 0.04535017221584386
['AQI', 'PM2.5', 'NO2', 'O3', 'CO'] 0.044489092996555686
['AQI', 'PM2.5', 'PM10', 'O3', 'CO'] 0.05381745120551091
['AQI', 'PM2.5', 'PM10', 'NO2', 'CO'] 0.05109070034443169
['PM2.5', 'PM10', 'NO2', 'O3', 'CO'] 0.0472158438576349
['AQI', 'PM2.5', 'PM10', 'SO2', 'CO'] 0.011481056257175661
['AQI', 'PM2.5', 'PM10', 'SO2', 'NO2', 'O3'] 0.016073478760045924
['AQI', 'PM2.5', 'PM10', 'NO2', 'O3', 'CO'] 0.04247990815154994
```

Figure 6. Correlation results for other time periods

Finally, it can be concluded that AQI is closely related to PM2.5, PM10, NO2, O3.

4. Conclusion

This paper analyzes air quality test data and explores the correlation between air pollutants and AQI using Apriori algorithm to identify the most relevant influencing factors for the purpose of reducing air pollution. According to the results of the study, it is necessary to suppress the sources of PM2.5, PM10, NO2 and O3 generation in order to protect the environment.

References

[1] B. Minaei-Bidgoli, R. Barmaki, M. Nasiri.Mining numerical association rules via multi-objective genetic algorithms [J]. Information Sciences.2013, 233: 15-24.

[2] Lilia Tightiz, Morteza Azimi Nasab, Hyosik Yang, Abdoljalil Addeh.An intelligent system based on optimized ANFIS and association rules for power transformer fault diagnosis[J].ISA Transactions. 2020, 103: 63-74.

[3] Mohammad Dehghani, Zahra Yazdanparast,Discovering the symptom patterns of COVID-19 from recovered and deceased patients using Apriori association rule mining, Informatics in Medicine Unlocked, Volume 422023,101351,ISSN 2352-9148

[4] Seth A. Horn, Purnendu K. Dasgupta,The Air Quality Index (AQI) in historical and analytical perspective a tutorial review,Talanta,Volume 267,2024,125260

[5] AGRAWAL R,IMIELINSKI T,SWAMI A.Mining associa tion rules between sets of items in large databases[C]// Proceedings of the 1993 ACM SIGMOD International Con ference on Management of Data,Washington,May 26-28, 1993.New York:ACM,1993:207-216.

[6] HAN J W,PEI J,YIN Y W,et al.Mining frequent patterns without candidate generation:a frequent-pattern tree app roach[J].Data Mining&Knowledge Discovery,2004,8(1):53-87.

[7] ZAKI M J,PARTHASARATHY S,OGIHARA M,et al.New algorithms for fast discovery of association rules[C]//Pro ceedings of the 3rd International Conference on Knowledge Discovery and Data Mining,Newport Beach,Aug 14-17, 1997.Palo Alto:AAAI,1997:283-286.

[8] Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Proc. 20th int. conf. very large data bases, VLDB. vol. 1215; 1994. p. 487–99. Santiago, Chile.

[9] Cheng C-W, Wang MD. Healthcare data mining, association rule mining, and applications. Health Inf Data Anal: Methods and Examples 2017:201–10.

[10] Rai VK, Chakraborty S, Chakraborty S. Association rule mining for prediction of COVID-19. Decision Making: Appl Manag Eng 2023;6(1):365–78.

[11] Ilbeigipour S, Albadvi A. Supervised learning of COVID-19 patients' characteristics to discover symptom patterns and improve patient outcome prediction. Inform Med Unlocked 2022;30:100933.