Relationship Analysis of AQI and Air Pollutants with Association Rule

Junlin Liao [1, 2, *]

[1]Department of School of Mathematics, South China Normal University, Guangzhou, China

[2]School of Mathematics, South China Normal University, Guangzhou, China

* Corresponding Author Email: 20202231006@m.scnu.edu.cn

Abstract. To analyze the main causes of bad air quality in recent years and the relationships embedded between air pollutants, this paper uses an improved Apriori algorithm to analyze the data mining of the correlations between overall air quality and air pollutants and between air pollutants in Guangzhou city's air quality data from 2014 to 2020.The data are divided according to AQI levels and air pollutant exceedance thresholds, and a Boolean type transaction dataset is constructed. Based on this, support, confidence and boosting thresholds are set to mine the transactional dataset for frequent item sets and obtain strong association rules. The results show that (1) The excess concentrations of PM10, CO, and NO_2 play a role in the excess concentrations of PM2.5. (2) In recent years, the main pollutant indicators of air pollution are O_3, NO_2, $PM_{2.5}$, PM_{10}, among which O_3 has the greatest influence. The strong correlation rules obtained through data mining have some significance to the detection and prevention of air pollution.

Keywords: AQI, Air pollutants, Association rules, improved Apriori algorithm.

1. Introduction

With the development of society and the deepening of communication in all areas of society, air quality conditions are closely related to all aspects of social life. Air quality is not only about public health, but also about social construction, and it has become a vital part of people's work to grasp the air quality situation and to monitor and prevent air pollution.

In recent years, AQI (Air Quality Index) has been widely used throughout the country due to the increasing awareness of environmental protection and the continuous development of environmental monitoring technology in various countries. The air quality level is determined by comparing it with predetermined air quality standards. As a composite index of multiple pollutants, AQI has a good effect in measuring the overall air quality situation.

In the monitoring and analysis of air quality data, a clearer understanding of the relationship between AQI and air pollutants and the relationship between each air pollutant will help our environmental management authorities take effective management countermeasures, which is of great significance for the prevention and control of air pollution. Therefore, this paper will analyze the relationship between AQI and air pollutants and the relationship among air pollutants by using data mining methods based on the daily AQI index and the corresponding air pollutant concentrations found.

2. Literature Review

With regard to the relationship between AQI and air pollutants and the relationship between air pollutants, many scholars in China have conducted research in recent years using correlation analysis and other measurement methods.In the study of the problem of the relationship between air pollutants, JinZhihao[1] used spss software to conduct correlation analysis and cluster calculation for each air pollutant for the monitoring data of ambient air pollutants in Hongkou District, Shanghai in 2019, and concluded that there is a strong correlation between each pollutant (NO_2, PM10, PM2.5 and CO all have significant positive correlations, while O_3 has negative correlations with other pollutants.) Considering that the time trend will have some influence on the relationship between each pollutant,

Wang Jin[2] introduced the detrended correlation coefficient to analyze the correlation between PM2.5 and other air pollutants in Urumqi to obtain the change of the correlation between air pollutants at different time points and different time scales.

For the issue of the relationship between AQI and air pollutants, the research methods of related scholars are different from the study of the relationship between air pollutants. For example, ZhaoYue [3]and others used pearson correlation coefficient to analyze the correlation between air pollutants and AQI during the heating period and non-heating period in Harbin, respectively, and concluded that air pollutants and AQI have significant correlation based on the correlation, and established multiple linear regression equation to further conclude that the air pollutant that plays a major role in the change of AQI is CO. Based on this, XuXin [4] in 2017 used the one-dimensional linear regression method to construct the one-dimensional regression equation of AQI and pollutant concentration as well as the least squares method to solve the coefficients of the one-dimensional regression equation, and then derived the influence of pollutant concentration on AQI. Besides, XuYunzhi [5] used the multiple linear regression method to construct the relationship equation between AQI and some pollutants based on the elimination of insignificant variables and sample outliers, and used the stepwise regression method to improve the problem of multicollinearity in the data.

Although the traditional methods of correlation and regression analysis are simple and easy to understand, their results are informative in exploring the relationships between air pollutants and with AQI. However, due to the limitations of their assumptions, correlation analysis and regression analysis are not easy to discover the implied a priori relationships between variables, and the correlation rule, also as a method to explore the relationships between variables, can solve this problem well. Therefore, in this paper, based on the discretization of AQI index and each air pollutant, we apply Apriori algorithm in association rules to carry out correlation analysis, and analyze the association relationship between AQI and air pollutants and air pollutants through the strong association rules derived from the mining, so as to provide reference significance for the monitoring and prevention of air pollution

3. Model Theory

3.1 Association rule mining method

As one of the main methods of current data mining, association rule mining [6] refers to exploring the correlation or causal structure among sets of variables by finding frequent occurrence sets in a given data set, and then discovering the correlation events hidden in the data that are not easily detected or even contrary to human consciousness.

The common association rule evaluation criteria are support, confidence and elevation, and their specific definitions and related calculation formulas are shown in Table 1:

Table 1. Definition and calculation formula of association rule evaluation criteria

Indicator Name	Explanation	Calculation formula
Support($X_i \rightarrow X_j$)	The number of occurrences of several associated data in the dataset as a proportion of the total dataset	$Support(X_i \rightarrow X_j)=P(X_i \cup X_j)$
Confident($X_i \rightarrow X_j$)	The probability that a data X_j appears after another data X_i appears.	$Confident(Xi \rightarrow Xj)=\frac{Support(Xi \rightarrow Xj)}{P(Xi)}$
Life(Xi→Xj)	The probability of containing data Xi at the same time under the condition that it contains data Xj.	$Life(Xi \rightarrow Xj)=\frac{Confident(Xi \rightarrow Xj)}{P(Xj)}$

Association rule mining was first used to study the relationship between "beer and diapers" in Wal-Mart supermarkets in the United States [7]. In recent years, with the increasing innovation ability of

people, association rules have been widely used in food safety [8], biological information [9], financial transactions [10] and other fields.

Common association rule mining algorithms include Apriori algorithm, FP-growth algorithm, XFP-tree algorithm. Considering that compared with other algorithms, Apriori algorithm is simple and efficient and has a better explanation of association rules, we use Apriori to develop association rule mining in this paper.

3.2 Association rule mining method

Apriori algorithm [11] [12] is an algorithm proposed by R. Agrawal and R. Srikant to mine the association relationship between data. The algorithm uses a layer-by-layer search iterative approach, first generating the candidate set Ck, then calculating the support of all k-itemsets, if the support meets the threshold requirement of greater than the minimum support, it becomes the frequent k-itemset L_k. On the basis of the candidate set C_{k+1} generated by the frequent k-itemset L_k, the frequent k+1 itemset L_{k+1} is determined by the threshold judgment of the support of the candidate set, and so on. until the next frequent itemset is not found. Different association rules are composed from the items in each frequent set, and the confidence degree of each rule is calculated separately, and if the confidence degree is greater than the minimum confidence degree threshold, the association rule is generated [13]. Therefore, the rules that satisfy both the minimum support and minimum confidence requirements become valid rules.

Although Apriori algorithm is simple and efficient and it has been widely used in association rule-related research, Apriori is subject to several problems in association rule operations as follows:

1. Low efficiency of candidate set generation: The Apriori Algorithm searches for frequent item sets by generating candidate sets, but the process of candidate set generation is very time-consuming on large-scale datasets. Since the number of candidate itemsets grows exponentially with the size of the itemset, this leads to a decrease in the efficiency of the algorithm.

2. Imperfect criteria for rule evaluation and selection: Apriori algorithm usually uses support and confidence as criteria for rule evaluation and selection, but these two metrics cannot fully capture the relevance and importance of the rules, and some of the association rules it generates may be contradictory. For example, suppose a total of 100 data, A appears 80 times, B appears 60 times, and A and B appear together 40 times. Then the probability of A and B appearing together is 40%. The probability of A appearing in the presence of B is 67%, while the probability of A appearing without B is 80%, which indicates that the presence of B decreases the probability of A so that A and B are not correlated.

To address the above issues, the following improvements are made to the Apriori algorithm in this paper [14]:

First, for the problem of low efficiency of candidate itemset generation, this paper develops an improvement of the algorithm structure, specifically: the elements in the frequent 1 itemset L_1 of the original algorithm are combined in two connections to obtain the candidate set C_2, and in order to obtain the frequent 2 itemset, all elements in the original database need to be scanned for the second time to find the support for all possible itemsets in the candidate set C_2. However, considering that there are some elements (samples) in the original database that are useless and do not contain any frequent itemsets. Therefore, we change the second scan of the database to a scan of the elements in the frequent 1 item set L_1 (the order of magnitude of the frequent 1 item set L_1 is much less than that of the original database). The candidate set C_2 is obtained by connecting the elements in L_1, and the support degree is calculated for the combination of elements in the candidate set C_2, which is again compared with the preset minimum support degree and pruned to obtain the frequent 2-item set L_2, and so on in a cyclic operation, in order to obtain the frequent K+1-item set, only the support degree is calculated for the elements in the previous candidate set C_K.

Second, for the problem of imperfect rule evaluation and selection criteria, this paper introduces the "elevation" indicator to further screen the association rules. The elevation indicator reflects how much the probability of occurrence of A has an impact on the occurrence of B. By determining the

elevation threshold, the problem of contradictory association rules can be effectively avoided. According to the definition of elevation, the threshold of elevation is 1.2:

Overall, the Apriori algorithm is divided into two steps, the first step generates the candidate set and the second step generates the association rules. Its algorithm flowchart is shown in Figure 1:

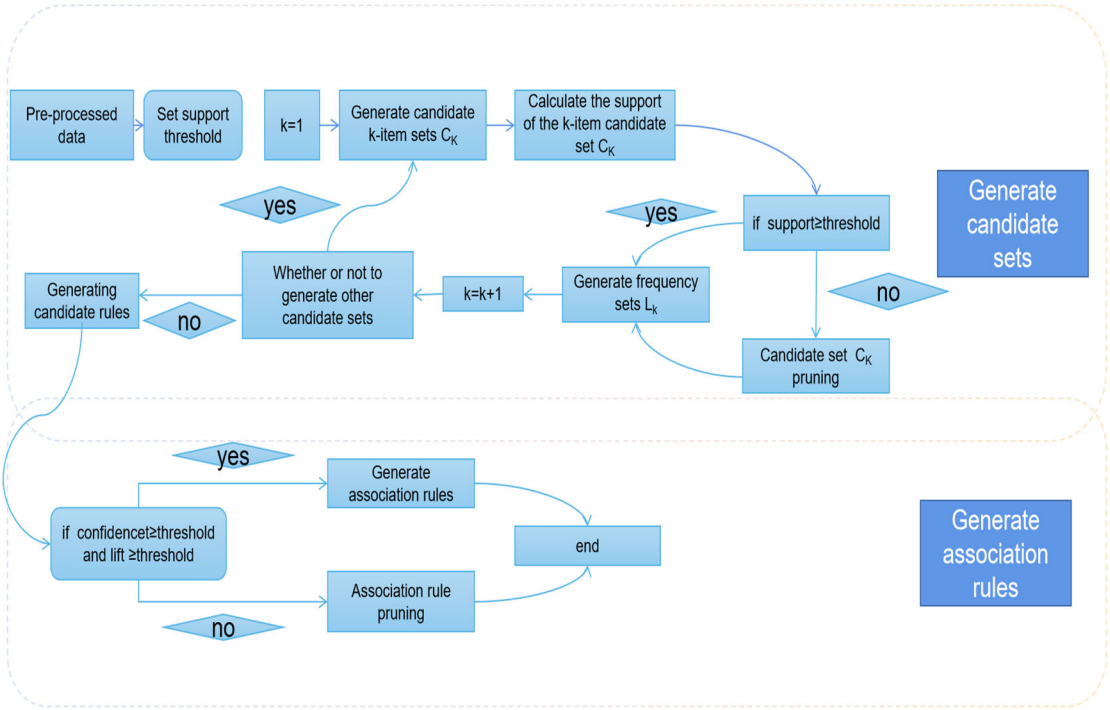


Figure 1. improved Apriori algorithm flowchart

4. Data Collecting & Processing

4.1 Data collecting

The data used in this paper were obtained from the State Environmental Protection Administration and the China Air Quality Online Monitoring and Analysis Platform, including the daily AQI, the 8h average O_3 concentration, and the daily average concentrations of five pollutants corresponding to AQI such as PM2.5 and PM10 in Guangzhou City from January 1, 2014 to June 30, 2020.

4.2 Data processing

4.2.1 Data Detection and Cleaning

The sample duplicates and missing values were retrieved by the dropna function and drop_duplicates function in the pandas library in python, and the data were detected as outliers according to the 3sigma principle. It was found that there were no obvious missing values, outliers and duplicates in this dataset, so there was no need to clean the data in this paper.

4.2.2 Data Type Conversion

Since the data of each variable in the dataset are of continuous type, they do not conform to the data types to which the association rules apply. Therefore, in this paper, continuous variables in the dataset are converted into Boolean variables based on the reference AQI class classification criteria [15] and ambient air quality criteria [16], and the specific transformation forms are shown in Table 2:

Table 2. Data type conversion of AQI and air pollutants

Variable name before conversion	Variable name after conversion	Transformation method
AQI	Appearance of air pollution	"AQI>100"="1", "AQI<100"="0"
$C_{O_3,8}$	O_3 (concentration exceeds standard)	"$C_{O_3,8}$>160"="1", "$C_{O_3,8}$<160"="0"
$C_{NO_2,24}$	NO_2 (concentration exceeds standard)	"$C_{NO_2,24}$>80"="1", "$C_{NO_2,24}$<80"="0"
$C_{SO_2,24}$	SO_2 (concentration exceeds standard)	"$C_{SO_2,24}$>50"="1", "$C_{SO_2,24}$<50"="0"
$C_{CO,24}$	CO (concentration exceeds standard)	"$C_{CO,24}$>1.5"="1", "$C_{CO,24}$<1.5"="0"
$C_{PM2.5,24}$	PM2.5 (concentration exceeds standard)	"$C_{PM2.5,24}$>55"="1","$C_{PM2.5,24}$<55"="0"
$C_{PM10,24}$	PM10 (concentration exceeds standard)	"$C_{PM10,24}$>100"="1","$C_{PM10,24}$<100"="0"

5. Model Establishment and Resolution

5.1 Analysis of association rules for air pollutants

The minimum support, minimum confidence, and minimum lift are set to 0.02,0.8,1.2, and the association rules derived by applying the improved Apriori algorithm are shown in Table 3:

Table 3. Correlation between different air pollutants

	Association Rules	Sup	Con	Lift
1	{'PM10'}→{'PM2.5'}	0.0754	0.9521	6.309
2	{'CO'}→{'PM2.5'}	0.0227	0.8710	5.7707
3	{'NO_2'}→{'PM2.5'}	0.0527	0.8621	5.7118
4	{'PM2.5', 'NO_2'}→{'PM10'}	0.0442	0.9545	6.3245
5	{'PM10', 'NO_2'}→{'PM2.5'}	0.0442	0.9545	6.3245

According to the association rules in Table 3, we can derive the association relationships between different air pollutants, which can be summarized as the following three points:

(1) The occurrence of 'PM10 concentration exceeds standard', 'CO concentration exceeds standard', and {'NO2 concentration exceeds standard'} events play a certain role in promoting the occurrence of {'PM2.5 concentration exceeds standard'} events, where {'PM10 concentration exceeds standard'} has the highest degree of promoting role in the occurrence of {'PM2.5 concentration exceeds standard'} events.

(2) The simultaneous occurrence of two pollutant concentration exceedance events will make {'PM2.5 concentration exceeds standard'} events more likely than PM10 and CO single pollutant concentration exceedance events.

(3) The occurrence of {PM2.5, NO2 concentration exceeds standard} events contributes to the occurrence of {'PM10 concentration exceeds standard'} events.

In summary, there is a strong positive correlation between "PM10 concentration exceeds standard", "CO concentration exceeds standard" and "NO2 concentration exceeds standard" and "PM2.5 concentration exceeds standard", which means that the exceedance of these pollutants may lead to the occurrence of "PM2.5 concentration exceeds standard". Such an analysis can help environmental authorities and the public to better understand the relationship between different pollutants and to take appropriate measures to improve air quality.

5.2 Analysis of association rules between air pollutants and AQI

The minimum support, minimum confidence, and minimum lift are set to 0.05,0.8,1.2, and the association rules derived by applying the improved Apriori algorithm are shown in Table 4:

Table 4. association rules between air pollutants and AQI

	Association Rules	Sup	Con	Lift
1	{'O_3 '}→{'Appearance of air pollution'}	0.104	0.980	5.915
2	{'NO_2 '}→{'Appearance of air pollution'}	0.05	0.896	5.411
3	{'PM2.5', 'PM10 '}→{'Appearance of air pollution'}	0.06	0.804	4.855

From Table 4, we can conclude that air pollution is mainly caused by exceeding the concentrations of four air pollutants, O3, NO2, PM2.5, and PM10, with O_3 posing the greatest threat to air pollution. Although O_3 plays a protective role in the atmosphere, high concentrations of O_3 can have a negative impact on both human health and the environment. Therefore, measures such as reducing tailpipe emissions and using clean energy should be taken to reduce the concentration of O_3 and the threat of O_3 to air pollution.

6. Conclusion

In this paper, data mining related techniques are applied to the correlation analysis between air pollution and air pollutants. Firstly, the variables measuring air pollution and air pollutants are transformed from continuous variables to Boolean variables through data processing based on the meaning of AQI values and the definition of air pollutant emission concentration standards. Based on this, the improved Apriori algorithm in data mining association rules is used to derive strong association rules between air pollutants and between air pollutants and overall air conditions, which can be used as a reference for further air pollution detection and prevention and control.

Compared with the traditional correlation analysis and regression analysis, the association rule analysis can quickly and intuitively explore the implicit relationship between air pollutants and the overall air conditions and air pollutants. Furthermore,The association rule algorithm can be introduced to explore the relationship between meteorological factors and air pollution to further reveal the mystery of air pollution.

References

[1] Jin Zhihao. Correlation analysis of regional ambient air pollutants[J]. Green Technology,2022,24(24): 226-228. DOI: 10.16663/j.cnki.lskj.2022.24.041.

[2] Wang J, Qiao YH, Zhang HG, et al. Analysis of the interrelationship number between the going trend of PM2.5 and other air pollutants in Urumqi[J]. Mathematical practice and understanding,2022,52(11):77-85.

[3] Zhang YJ, Zheng Y, Jiang B et al. Study on the correlation analysis of AQI and air pollutants in Harbin City[J]. Environmental Science and Management,2017,42(02):74-76.

[4] Xu Xin. Research on air quality index in Hangzhou based on linear regression analysis[J]. China New Communication,2019,21(01):239-241.

[5] Xu Yunzhi, Fan Yingying, Yao Yufei, Sun Hongwen, Yang Yi, Li Pengcheng. Application of multiple linear regression method in air quality index AQI analysis[J]. Coal mine electromechanics,2019,40(05):57-62.DOI:10.16545/j.cnki.cmet.2019.05.015.

[6] Sun Xiange. The application and impact of data mining technology in the field of library intelligence [J]. Intelligence Exploration,2009,No.(08):85-87.

[7] Jian Xiangdong. Research on the application of association rules in teachingevaluation[D]. Yan'an University, 2022. DOI:10.27438/d.cnki.gyadu.2022.000383.

[8] Xue W.B., Wang S.D., Li M.L. et al. Association rule mining of multi-component mycotoxin contamination in wheat based on Apriori algorithm[J]. Chinese Journal of Food Hygiene,2022,34(03):451-458.DOI:10.13590/j.cjfh.2022.03.010.

[9] Zhuang W B,Ma J C et al. Data mining-based study of anti-influenza virus herbal medicine dosing patterns[J/OL]. Pharmacy Today:1-5 [2023-04-11]. http://kns.cnki.net/kcms/detail/44.1650.R.20230406.1814.004.html.

[10] Peng Hui, Liu Xinyu. A study of industry rotation phenomenon in Chinese stock market based on correlation rules[J]. Journal of Beijing University of Posts and Telecommunications (Social Science Edition),2016,18(01):66-71.

[11] Wang Zengping, Zhao Bing, Li Zekun, Sun Yi, Li Bin. Analysis of factors influencing the quality of energy meter verification based on Apriori association rules[J]. Electrical Measurement and Instrumentation,2017,54(13):85-92.

[12] Yang Fang. Apriori association rule-based evaluation method of smart energy meter failure rate level in typical environment[D]. Hunan University,2020.DOI:10.27135

[13] Wu Wenbo. Data mining technology applying Apriori association rule algorithm for mining e-commerce potential customers [D]. Hangzhou: Zhejiang University of Technology, 2012.

[14] WangMY, Ding H, Xu R. Correlation factor analysis of ambient air NO2 concentration variation based on improved Apriori algorithm[J]. Journal of Tropical Meteorology,2022,38(06):890-900.DOI:10.16032/j.issn.1004-4965.2022.080.

[15] Zhao Y. Air quality index prediction based on ARIMA and SVR combined model[D]. Tianjin University of Commerce,2019.

[16] Ambient Air Quality Standards (GB 3095-2012). http://www.wxbh.gov.cn/