Contents lists available at YXpublications

International Journal of Applied Mathematics in Control Engineering

Journal homepage: http://www.ijamce.com

Research on Correlation Analysis of Air Quality Elements Based on Spark

Qingyuan Wang[a], Long Zhang[b], Xiaoxuan Song[a], Honggang Li[a], Wenfeng Qin[a], Qingjie Zhao[b, *], Bingxin Niu[c]

[a] *Hebei Xiongan Rongwu Expressway Co., Ltd., Hebei, 071700, China*

[b] *Hebei Provincial Commnications Planning, Design and Research Institute Co., Ltd, Hebei, 050000, China*

[c] *Hebei University of Technology, Tianjin, 300401, China*

ARTICLE INFO

Article history:
Received 2 November 2022
Accepted 20 November 2022
Available online 5 December 2022

Keywords:
Air quality
Association rules
Distributed computing
Air environment treatment
Spark

ABSTRACT

In order to find out the correlations in the huge amounts of air quality data and mine the valuable hidden information from the data, the algorithm of association rule was used to analyze the three-month months air quality monitoring data of a city in this paper. Firstly, clustering algorithm was used to discretize the collected air quality data. Then, the discrete data was input to the improved the association rule mining algorithm-BCLARM algorithm for Spark distributed computing based on load balancing, so that it is super-efficient to mine the association rules of the data. Finally, the strong association rules mined from the data are analyzed and explained to provide a theoretical basis for air environment governance.

1. Introduction

In recent years, with the rapid development of our country's economic construction and social productivity, air pollution has become an important topic of special concern to the public and the government. With the development of science and technology, the measurement technology and instruments for air pollutant emissions have been continuously updated. China has set up more than 5,000 air quality monitoring sites at the national, provincial, municipal and county levels. A large number of air pollution and meteorological related data. However, the air pollution indicator system is complex and the amount of data is huge. Assuming that each monitoring site can record air quality data every hour, millions of data will be generated within a week, and the amount of data generated every year is even greater. Therefore, relevant theoretical methods of data mining can be used to analyze a large number of data on changes in air quality and meteorological elements to provide information for urban environmental capacity verification, industrial structure layout, urban planning and construction, resource development and utilization, and improvement of human living environment. Propose a more reasonable ecological environmental protection plan, and provide data support and guarantee (Qin et al., 2015; He et al., 2018; Song et al., 2018). Nowadays, more and more scholars use the correlation theory method in data mining to explore the correlation between air quality indicators and air pollution, meteorology, diseases and other phenomena, and use association rule technology to conduct in-depth research in the field of meteorology (Xie et al., 2015; Cagliero et al., 2016; Sun, L. et al., 2019; Han et al., 2021). In 2018 , Wang Zhe et al. used the association rule between the number of visits for COPD and meteorological conditions and air quality (Han et al.,2021). In 2015, Shan et al. applied association rules to the correlation analysis between the Beijing-Tianjin-Hebei region and particulate matter concentration, which provided help for the prevention and control of PM (Qin et al., 2015). In 2018 , Liang Songyue et al. used spatial interpolation, statistical analysis and other methods to study the spatiotemporal distribution of $PM_{2.5}$ mass concentration in the Beijing-Tianjin-Hebei region and its correlation with PM_{10}, sulfur dioxide, nitrogen dioxide, ozone, and carbon monoxide (Liang et al., 2018); In 2020 , Asiguri Niyaz and others used Spearman to analyze the relationship between air pollutants and the correlation between air pollutant concentrations and children's respiratory diseases, and concluded that the concentration of major pollutants in the atmosphere is closely related to the common air pollutants in children. Respiratory diseases related (Asigul et al., 2020). In 2016, Deng et al. proposed the DFIN (Qin et al., 2015) algorithm, using the new DiffNodeset structure to improve the mining efficiency of frequent itemsets; in 2018, Gu Junhua (Deng,

* Corresponding author.
E-mail addresses: 892774766@qq.com (Q. Zhao)
doi:

2016) and others implemented the optimization of the FP-Growth algorithm on the Spark platform. The balanced division of the data set is fully considered, which greatly improves the parallelism of the algorithm. In 2019, Xu Liangliang et al. used the Apriori algorithm to mine the association rules between power load data and meteorological factors to explore the degree of influence of meteorological factors on power load (Gu et al, 2018). In 2020, He Liwei et al. (2020) established a prediction model through the multiple regression method; at the same time, the sensitivity of meteorological factors was introduced to analyze the contribution of meteorological factors in power grid load forecasting, and quantitatively analyze the impact of meteorological factors on power grid load. Therefore, association rules algorithm can be used to mine the correlation between air quality indicators and provide decision support for the formulation of air pollution control policies.

This paper proposes an improved Spark-based association rule algorithm for mining massive air quality data and analyzing the implied associations. This algorithm implements an improved association rule algorithm BCLFARM algorithm proposed in the previous research of the research group on the Spark distributed computing platform, and proposes a grouping strategy to achieve load balancing. On the basis of improving the mining efficiency of traditional association rule algorithms. The algorithm has better parallelism and can complete the data mining of huge air quality data in an acceptable time.

2. System structure of air quality index correlation mining

In order to explore the correlation between air quality indicators, it is first necessary to collect data. The paper uses relevant monitoring instruments to collect air quality indicators and meteorological elements monitoring data per hour in 9 districts and counties in a city for 90 days. The recorded air quality indicator data includes $PM_{2.5}$, PM_{10}, $S0_2$, NO_2, etc. Meteorological elements include temperature, humidity, air pressure, etc. Then the data is processed, including data cleaning, integration, normalization and discretization, so that the data format can meet the input requirements of the association rule algorithm (Xu et al., 2019; Deng, 2016).

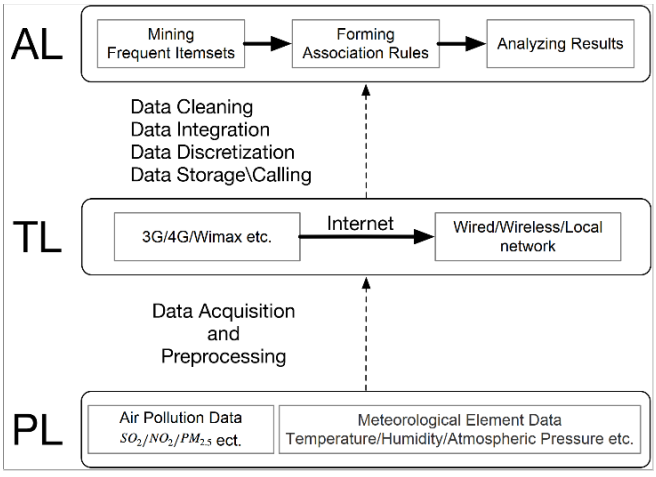

Figure 1 The flow chart of association relationship mining

Finally, input the sorted data into the association rule algorithm in a certain format, set the minimum support and minimum confidence, explore the association rules, analyze and evaluate the generated association rules, and filter out valuable hidden association information to provide certain decision-making support for the relevant departments to formulate environmental governance policies (Bui et al.,2018). The air quality index correlation mining is divided into three levels, and the specific mining process is shown in Figure 1.

Physical layer: This layer needs to use monitoring equipment to observe changes in air quality and meteorological factors in real time, and record relevant data as a data source for data mining.

Transmission layer: This layer needs to use a limited network or wireless network to transmit the real-time monitoring data, and send the data to the main monitoring station. It is necessary to ensure that the network is unblocked as much as possible to ensure the continuity of the data.

Application layer: After cleaning, integrating and discretizing the data, perform association rule mining on the data, and analyze the final mining results to obtain implicit valuable information.

3. Air quality data collection and processing

3.1 Data collection and preprocessing

By installing monitoring equipment in different areas of the city, air quality data and meteorological data are collected, and the hourly data of each monitoring point is recorded as a piece of raw data, transmitted through the wireless network, and all data are summarized and recorded in the total monitoring system. Stations are stored in excel form by month.

Since the original data collected during the data collection process often has problems such as data missing, data redundancy, noise interference, etc., it is necessary to perform a series of effective data processing on the original data, such as cleaning the incomplete data in the original data. , to correct or clear inconsistent data.

(1) Data cleaning

In the original data, data transmission may fail due to network congestion and other reasons, resulting in data that is discontinuous in time for a monitoring device. Such empty data will have a certain impact on the final mining result. Therefore, it is necessary to delete data missing and abnormal data in the original data to prevent these data from affecting the final data mining results.

(2) Data integration

Integrate the data collected by multiple data monitoring instruments, so that the data are sorted in a certain time and space order. In this paper, the data is sorted according to the daily and hourly data of each monitoring point as an order, so that the messy data becomes orderly.

(3) Normalization

After the data is organized in an orderly manner, it is necessary to normalize the data of different standards, that is, the dimensional expression is transformed into a dimensionless expression and becomes a scalar. This way of changing the absolute value of the system value into a relative value can simplify the calculation and make the presentation of the data more intuitive.

3.2 Data discretization

Air pollution and meteorological element data are both continuous data. In order to use the frequent itemset mining algorithm to analyze the data, they must be discretized first. In this paper, the method of clustering discretization is used to discretize the data.

In addition to meteorological data, the experiment also added primary pollutants, pollution source types and geographical location information of monitoring sites as data mining items. These items can be used to analyze the reasons affecting air quality from various

perspectives, such as environmental background types, geographical perspectives Etc., analyze the causes of air pollution through thinking expansion. All other mining items are shown in Table 1.

Table 1 Other mining data items table

Mining items	Mining item value				
primary pollutant	$PM_{2.5}$	PM_{10}	O_3	NO_2	CO
Type of pollution source	Construction site	sensitive point	main road	key enterpri-ses	traffic road
	School	dining area	Hospital	urban village	other
Districts and counties	County A	County B	County C	County D	County E
	County F	County G	County H	County I	

The continuous attributes are clustered by the K-Means clustering algorithm, and then the k clusters after the clustering are processed to obtain the classification value corresponding to each cluster, and the discretization of the continuous data is completed. For example, using all PM_{10} data for cluster division, the continuous data of PM_{10} is divided into "0-6 2 ", "63-14 7 ", "148-25 7 ", "258-464" and ">464" Five levels, each level is regarded as a discrete value after the discrete division of a cluster. Among them, for the division of wind direction, the method of clustering discrete is not used, but the wind direction is divided into four directions, and each direction is regarded as a discrete value. The results of the discrete meteorological data of all continuous attributes are shown in Table 2.

Table 2 Discretized data grouping table

Meteorological data	Level one	Level two	Level three	Level four	Level five
PM_{10}	0-62	63-147	148-257	258-464	>464
	M1	M2	M3	M4	M5
$PM_{2.5}$	0-36	37-84	85-145	146-253	>253
	P1	P2	P3	P4	P5
SO_2	0-12	13-36	37-65	66-123	>123
	S1	S2	S3	S4	S5
NO_2	0-19	20-54	55-90	91-802	>802
	N1	N2	N3	N4	N5
O_3	0-18	19-55	56-97	98-168	>168
	O1	O2	O3	O4	O5
CO	0-0.6	0.7-1.9	2.0-3.8	3.9-7.9	>7.9
	C1	C2	C3	C4	C5
temperature	-1.6-0	1-6	7-16	17-49	>49
	T1	T2	T3	T4	T5
humidity	0-32	33-46	47-62	63-78	>78
	H1	H2	H3	H4	H5
wind direction	0-89	90-179	180-269	270-360	
	WD1	WD2	WD3	WD4	
wind speed	0-0.9	1-1.9	2-3	>3	
	WS1	WS2	WS3	WS4	
air pressure	900-999	1000-1004	1005-1010	>1010	
	P1	P2	P3	P4	

After completing the data processing and discretization, the available data is stored in the MySQL database to ensure that the data can be quickly called when mining multiple frequent itemsets with different support thresholds.

A well resolved (spatially and temporally) and highly accurate direct numerical simulation (DNS) tool has been developed to understand the fundamental difference in the hydrodynamics of flow over two-dimensional and three-dimensional ripples in a channel geometry and its implications on sediment transport. As a first step, in this paper we focus on the steady flow over the ripples.

4. Association Rules Mining Algorithms

4.1 Basic Concepts of Association Rules

Let set $I = \{ i_1, i_2, \ldots, i_n \}$ be the set of general itemsets. DB = $\{ T_1, T_2, \ldots, T_m \}$ is the set of all transactions in the transactional database, |DB| = m , each transaction T_i ($i \in [1,N]$) has a unique identifier, and each transaction T_i is contained in I.ie $\forall T_i \in T$, $\exists T_i \subseteq I$.

(1) Itemset: A set containing 0 or more items. If the set $X=\{ x_1, x_2, \ldots, x_k \}$ is a subset of $X \subseteq I$, *that is, X is* called an itemset.

(2) k -itemsets: itemsets containing k items are called k -itemsets.

(3) Support degree: Given a transaction database DB as shown in Table 1 and a set of sets $A = \{a_1, a_2, \ldots, a_k\}$, A is a subset of $A \subseteq I$, *that is, A* is called Itemsets in *I*. If an itemset *A* contains k items, then *A* is called a k -itemset. The support count of *A is the number of transactions* that contain item set *A in the* DB, denoted as *support_count* (*A*). *A* 's support is defined as the ratio of *A* 's support count to the total number of transactions, denoted as *support* (*A*) , and the formula is expressed as:

$$\boldsymbol{support(A) = \frac{support\_count(A)}{|DB|} = \frac{support\_count(A)}{m}} \quad (1)$$

(4) Confidence: $X \to Y$ An expression of the form can be used to represent association rules, in which $X \cap Y = \varnothing$ the association strength can be measured by confidence. Confidence indicates the frequency of occurrence of another itemset *Y contained in X after* one itemset *X* appears. Denoted as *confidence(X→Y)*, its formula is expressed as:

$$\boldsymbol{confidence(X \to Y) = \frac{support(X \cup Y)}{support(X)} = \frac{support\_count(X \cup Y)}{support\_count(X)}} \quad (2)$$

(5) Frequent itemsets: If the support of the itemset *A* is greater than the minimum support *minSup* ($\in$ [0,1]), that is, it exists

$support(A) \geq minsup$, then A is a frequent itemset; if the length of A is k, then A is called Frequent k -itemsets.

(6) Strong association rule: The proportion that transaction database has both itemsets X and Y is signed by $support(X \cup Y)$. $support\_count(X \cup Y)$ is the number of transactions which have both itemsets. Set the minimum confidence as $minConf$ ($\in[0,1]$), and if $confidence(X \rightarrow Y) \geq minconf$, it means that $X \rightarrow Y$ is a strong association rule (Sadat Y. K. et al, 2015 and Yin, X. et al, 2019).

4.2 Algorithm Description

In order to improve the mining efficiency of frequent itemsets in association rule algorithm, a frequent itemset mining algorithm based on Bitmap-code List (BC-List) proposed by the research group (BC-List Frequent Itemsets Mining, BCLARM) was used.), which is used to solve the problems of complex construction and insufficient mining efficiency in the classical frequent itemset mining algorithm FP-Growth algorithm. The algorithm first uses the node encoding model based on bitmap representation to generate a bitmap tree (BC-tree), and uses the node information of the BC-tree as the data structure to quickly obtain the node set of the BC-List through bitwise operations, avoiding complex intersections. It improves the connection efficiency; secondly, by using superset equivalence and support count pruning strategy, the search space for mining frequent patterns is reduced. Experiments show that this algorithm has a faster mining speed than the FIN algorithm and the DFIN algorithm.

BCL ARM algorithm can be divided into the following steps:

(1) Scan the transaction database DB, sort the itemsets of each transaction in non-ascending order according to the support degree, obtain the frequent 1-itemsets F_1, and construct the BC-tree corresponding to the frequent 1-itemsets F_1.

(2) By scanning the BC-tree, the BC-List corresponding to the frequent 1-item set F_1 is obtained.

(3) In the search space represented by the set enumeration tree, use the superset equivalence and support count pruning strategies to mine frequent k-itemsets through frequent (k-1)-itemsets.

(4) Calculate the support of all k-itemsets.

In order to make the algorithm more efficient when processing massive data, the Spark distributed computing platform is used to implement the algorithm, the serial algorithm is parallelized, and the total computing time is reduced through multi-CPU parallel computing. A grouping optimization strategy based on load balancing is added to optimize the calculation method of the algorithm. Since the parallel grouping method of the Spark platform is very simple, the calculation amount of each node cannot be equal, which will cause the load imbalance problem that the overall calculation time is prolonged due to the excessive calculation amount of a single node, in order to solve this problem, first put forward the concept of frequent itemsets with the same base.

4.2.1 Homogeneous frequent itemsets

Same-base frequent itemsets: If there are frequent 2 - itemsets aa1, aa2, aa3, …aan, then these n frequent 2-itemsets are called the same base frequent 2-itemsets, and they have a common base a.

For a group of frequent itemsets with the same base based on a, all frequent k-itemsets based on a base can be obtained by connecting themselves within the group; and due to the nature of the connection, the frequent 2-itemsets of different bases are merged. There is no dependency, so parallel frequent itemset mining can be performed on each group of frequent 2-itemsets of different bases. Moreover, according to the example set enumeration tree in Figure 3.8, the connection probability of frequent 2-itemsets obtained by connecting frequent 1-itemsets with low support will decrease according to the decrease of support, so there is no need to consider the existence of isolated identical itemsets. The case of the base frequent itemsets.

For example, for two groups of different same-base frequent 2-itemsets F1 ={da,db,dc}, F2 ={ea, eb}, through their own intra-group connection, itemsets T1 ={da, db, dc, dab, dac, dbc, dabc}, T2 ={ea, eb, eab}, and there is no correlation between these two sets of itemsets obtained by connecting themselves, so parallel mining can be performed.

4.2.2 Packet optimization strategy based on load balancing

All frequent 1 -itemsets according to different bases can realize parallel frequent itemset mining of BCLARM algorithm, but due to the difference in the number of frequent itemsets between each group, it needs to be considered when dividing different groups into nodes. Otherwise, the overall computing time will be prolonged due to the excessive computing load of some nodes. When considering load balancing, the core idea is to try to make the time consumption of each computing node similar or the same, and to make the node with the longest calculation time the shortest time consumption, so as to make the overall calculation efficiency the highest. The load balancing optimization strategy objective is shown in Equation 3.

$$\min\left(\max_{1\leq n\leq N}\sum_{bclistk\in T_n} t_k\right) \tag{3}$$

where t_k represents the estimated computing time of each same-base grouping, $t_k = |bclist_k|$, bclist k is the total size of the BC-List of the same-base grouping K, that is, using $bclist_k$ as the approximate estimated time-consuming. The specific load balancing optimization strategy is as follows:

(1) First, estimate the time consumption of each frequent 2-itemset same-base grouping, that is, calculate $t_k = |bclist_k|$, and obtain the estimated time consumption of each grouping.

(2) Sort each group in descending order according to the estimated time T.

(3) Perform initial allocation to N nodes, that is, allocate the first N groups in descending order to N nodes. Calculate the estimated total time sumT for each group.

(4) The remaining groups are preferentially allocated to the node with the smallest sumT, and $sumT=sumT+T_k$, that is, sumT is updated after each allocation. Figure 2 shows an example of the grouping process.

4. 3 The overall process of the algorithm

The implementation of the BCLARM algorithm on the Spark platform is similar to the implementation of the parallel FP-Growth algorithm. First, it is necessary to obtain the BC-tree and the BC-List of frequent 1-itemsets and frequent 2-itemsets; then according to the grouping that satisfies load balancing. The strategy is to distribute the frequent 2-itemsets to different working nodes for calculation. After each working node obtains part of the frequent k-itemsets, they are merged to obtain all the frequent k-itemsets, and the association rules

are generated; Filter out strong association rules. The BCLARM algorithm is divided into four parts, as shown in Figure 3.

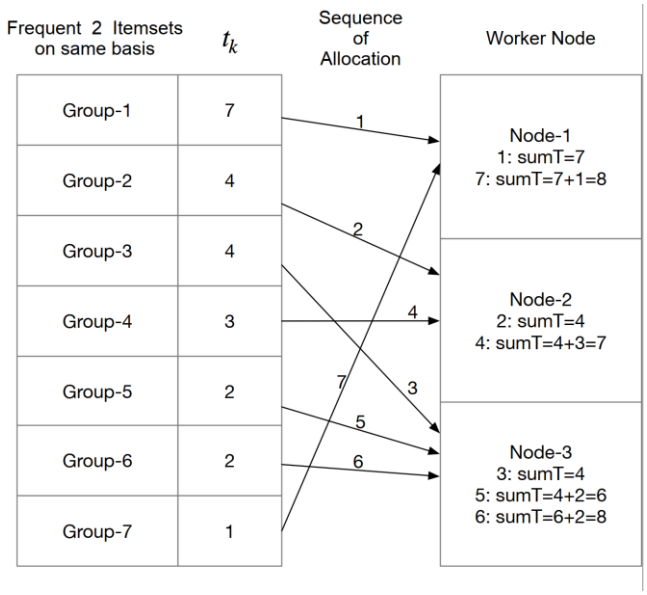

Figure 2 Grouping process under load balancing strategy

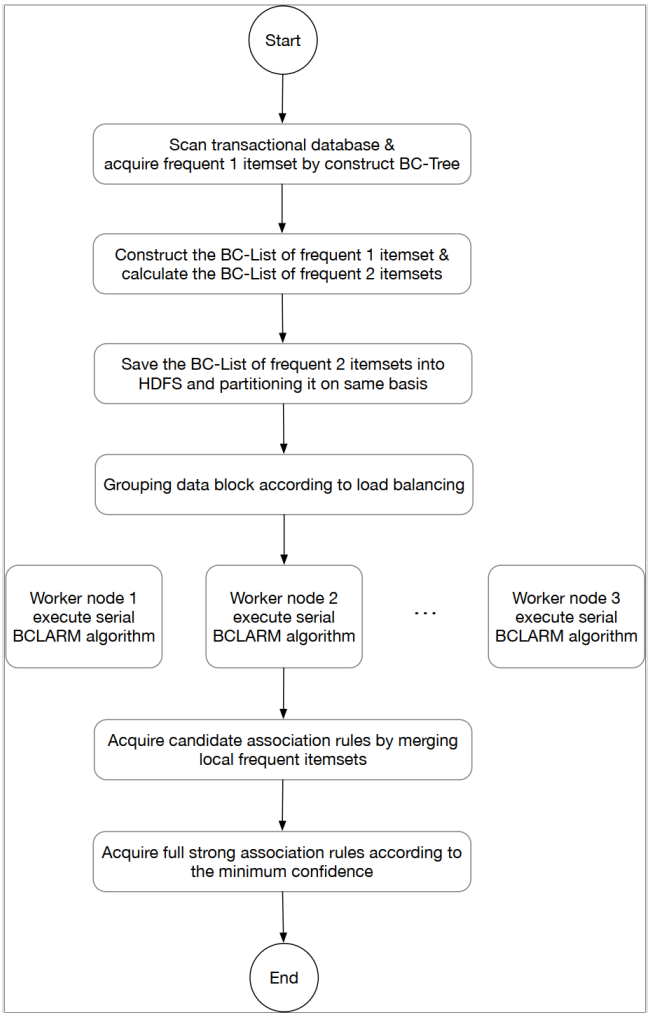

Figure 3 BCLARM algorithm flow

The steps of the BCLARM algorithm are described in detail as follows:

(1) Obtain the BC-List of frequent 2-itemsets

According to Algorithm 3.3 in Chapter 3, scan the entire transaction database, build a BC-tree, and obtain the BC-List of frequent 1-itemsets and all frequent 2-itemsets;

(2) Database division based on load balancing

Store the BC-List of all frequent 2-itemsets in a certain format in the HDFS file system (Hadoop distributed file system) , divide the frequent 2-itemsets of the same base into a data block, and divide the data through the above grouping strategy The blocks are evenly divided into each worker node;

(3) Obtain local frequent k-itemsets

Each worker node runs the serial BCLARM algorithm at the same time.

Perform frequent itemset mining on the data in the current working node, and obtain some frequent k-itemsets under the current working node;

(4) Combine calculation results and obtain association rules

The mining results of each worker node are merged to obtain global candidate association rules. According to the set minimum confidence, the strong association rules among all candidate association rules are mined.

5. Result analysis

5.1 Association rules mining

In order to improve the efficiency of the algorithm, the algorithm is designed and written in a distributed manner on the Spark platform. This experiment runs on a cluster consisting of 5 worker nodes, of which 1 worker node serves as both master and worker nodes, and the other 4 as a slave node. Each node has the same configuration and is connected by the same local area network. The specific hardware configuration is shown in Table 3.

In order to mine valuable association rules, after several experiments, set the minimum support to 5% and set the minimum confidence to 30% to mine the association rules of air quality data, it is not meaningful to remove some of the association rules. After retaining the results that meet the mining target, the final association rule results obtained by mining are shown in Table 4.

Table 3 Hardware configuration parameters

hardware	parameter
CPU	Intel(R) Xeon(R) CPU E5-2620 v4 @2.10GHz
Memory	32GB
hard disk	1TB
network	100mbps

Table 4 Association rules for air pollution elements

S/N	rule	Support	Confidence
1	M1=>P1	32.5%	96.8%
2	T1=>M1	10.3%	38.6%
3	H1=>M1	11.7%	58.3%
4	Key Enterprise => P1	15.6%	46.5%
5	H1=>M2	6.5%	31.9%
6	H2=>M2	8.4%	41.5%
7	H3=>M2	9.6%	37.3%

8	H4=>M2	8.3%	36.1%
9	Lubei District => M2	11.3%	38.3%
10	Construction site => M2	5.5%	43.8%
11	Main road => M2	6.8%	41.9%
12	Key Enterprise => M2	12.1%	31.1%
13	T4=>M3	6.1%	31.3%
14	Key Enterprise => S3	8.9%	56.7%
15	Key Enterprise => C3	8.8%	55.5%
16	H5=>PM2.5	9.1%	59.5%
17	Lubei District => PM2.5	12.4%	42.1%
18	Construction site => PM2.5	5.8%	46.5%
19	Main road => PM2.5	6.9%	42.6%
20	Key enterprises => PM2.5	12.2%	31.1%
21	T1=>M1 ∧ P1	7.9%	37.5%
22	N2 ∧ C2=>S2	7.2%	70.1%
23	S2 ∧ C2=>N2	7.2%	48.1%
24	S2 ∧ N2=>C2	7.2%	71.1%
25	S2 ∧ O2=>C2	6.2%	70.8%
26	M3 ∧ P3=>PM2.5	11.9%	90.7%
27	M3 ∧ P2=>PM10	11.8%	83.3%

5.2 Association rule analysis

First, convert the association rule codes in Table 4 into corresponding discrete values, and then analyze and interpret the generated association rules. The converted rules are shown in Table 5.

Table 5 Interpretation of association rules for air pollution elements

S/N	rule	S/N	rule
1	PM 10 (one) => PM 2.5 (one)	15	Key enterprises => CO (three)
2	temperature(one) => PM 10 (one)	16	Humidity (5) => Primary pollutant (PM 2.5)
3	temperature(one) => PM 10 (one)	17	Lubei District => Primary Pollutant (PM 2.5)
4	Key Enterprises => PM 10 (One)	18	Construction Site => Primary Pollutant (PM 2.5)
5	Temperature (one) => PM 10 (two)	19	Main Road => Primary Pollutant (PM 2.5)
6	temperature(two) => PM 10 (two)	20	Key Enterprises => Primary Pollutants (PM 2.5)
7	Temperature (three) => PM 10 (two)	21	temperature(one) => M1 ∧ P1
8	Temperature (four) => PM 10 (two)	22	NO (two) ∧ CO (two) => SO 2 (two)
9	Lubei District => PM 10 (Two)	23	SO 2 (two) ∧ CO (two) => NO (two)
10	Construction site => PM 10 (two)	24	SO 2 (two) ∧ NO (two) => CO (two)
11	Main Road => PM 10 (Two)	25	S O 2 (two) ∧ O 3 (two) => CO (two)
12	Key Enterprise => PM 10 (two)	26	PM 10 (three) ∧ PM 2.5 (three) => primary pollutant (PM 2.5)
13	temperature (four) => PM 10 (three)	27	PM 10 (three) ∧ PM 2.5 (two) => primary pollutant (PM 10)
14	Key Enterprise => SO 2 (three)		

In Table 5, the numbers in parentheses of each data item represent the division level of the item after clustering and discretization, for example, PM_{10} (one) represents the level one of PM_{10}. Among them, the confidence levels of the 26th and 27th rules are as high as 90.7% and 83.3%, which can prove that the correct information can be mined through the F-BCLFARM algorithm. According to Table 5 , the following information can be obtained from the analysis:

(1) In the four humidity levels, PM_{10} was slightly higher, but there was no obvious increase or decrease, that is, there was no obvious correlation between humidity increase and PM_{10}. The impact of the available humidity on PM_{10} may not have the positive and negative relationship shown; or in a small area, the PM_{10} concentration will not be affected by humidity when it reaches a certain level.

(2) The PM_{10} level in Zone A is slightly higher most of the time, and its main pollutant is $PM_{2.5}$. The analysis shows that the air pollution degree of Area A is relatively high, and it should be regarded as the key air environment control area.

(3) $PM_{2.5}$ in construction sites, main roads and key enterprises was slightly higher, and SO_2 and CO concentrations in key enterprises were higher. The analysis shows that the $PM_{2.5}$ of construction sites, main roads and key enterprises should be rectified mainly, and the harmful gases emitted by key enterprises should be rectified.

(4) NO_2, CO and SO_2 have obvious correlations and may influence each other. When carrying out air environmental treatment, attention should be paid to the coordinated treatment of the three polluting gases.

(5) When the temperature is low, the concentration of PM_{10} is low, and when the temperature increases, the concentration of PM_{10} also increases. There is a certain positive correlation between PM_{10} and temperature.

By analyzing the above information, it is found that the association rules obtained by mining are basically consistent with the actual situation, and some information that cannot be found intuitively can also be explained by objective facts. These potential rules discover the influence relationship between air pollution elements, provide decision makers with decision-making basis for air environmental governance, and also lay a foundation for air pollution prediction.

6. Conclution

In this paper, through the collection and processing of air quality index data in a city, discrete data that can be used for correlation

analysis of air quality indicators are obtained. Meaningful association rule information. After analyzing and explaining these association rules with high support and confidence, it can provide a certain theoretical basis for the formulation of air environmental governance policies. This method can process the collected mass air quality indicator data, and can use the discretization method to process different types of data. However, this method does not involve the temporal characteristics of air quality, so the next work plan is in the air quality time characteristic data was added to the correlation analysis to further improve the method and obtain more accurate correlation analysis results of air quality indicators.

Acknowledgements

This work would like to acknowledge financial support provided by scientific research project (the fourth batch) of KT12 section of the new line of Rongwu Expressway.

References

Qin, S., Feng,L., Chen,W., et al., 2015. Spatial-Temporal Analysis and Projection of Extreme Particulate Matter (PM10 and PM2.5) Levels Using Association Rules: A Case Study of the Jing-Jin-Ji Region. J. Atmospheric Environment 120,339–50.

He,W., Song,G., Liu,S., 2018. Urban PM2.5 Pollution Health Loss and Prevention Cost-Benefit Estimation: Taking Benxi City as an Example. J. Environmental Science.

Song,G., Guo,X., Yang,X., et al., 2018. ARIMA-SVM combined prediction of PM(2.5) concentration in Shenyang. J. China Environmental Science. 38(11),33-41.

Xie,C., Ma,M., Yu,X., 2015. Application of various neural networks in urban air quality prediction in western North China. J. Chinese Journal of Environmental Engineering. 9(12),361-365.

Cagliero,L., Cerquitelli,T., Chiusano,S., et al., 2016. Modeling Correlations among Air Pollution-Related Data through Generalized Association Rules. C. IEEE International Conference on Smart Computing (SMARTCOMP).

Sun, L., L. Sun, and Y. Li, 2019. *Development of Air Conditioning Fault Detection Technology in Existing Buildings.* International Journal of Applied Mathematics in Control Engineering. **2**(2): p. 170-175.

Han,G., Huang,Y., Jiang,N., et al., 2021.Research on meteorological data quality control based on improved Apriori algorithm. J. Electronic Test. (05), 63-64.

Wang,Z., Li,L., Wang,K., et al., 2018. Study on the relationship between the number of patients with chronic obstructive pulmonary disease and meteorological air conditions based on association rule analysis . J. China Digital Medicine.13(04),2-47.

Qin, S., Feng,L., Chen,W., et al., 2015. Spatial-Temporal Analysis and Projection of Extreme Particulate Matter (PM10 and PM2.5) Levels Using Association Rules: A Case Study of the Jing-Jin-Ji Region. J. Atmospheric Environment 120,339–50.

Liang,S., Lei,J., Zhang,A., 2018. Spatial-temporal distribution of PM_(2.5) mass concentration in Beijing-Tianjin-Hebei region and its correlation with five air quality indicators including PM_(10) . J. Jiangsu Science and Technology Information.35(22), 40-45.

Asigul,N., Zhang,M., Kurbanjan,K., et al., 2020. Talking about the main air pollution in Urumqi Study on the correlation between the concentration of the drug and the daily outpatient number of children with respiratory diseases. J. Xinjiang Medicine.50(02),166-168.

Qin, S., Feng,L., Chen,W., et al., 2015. Spatial-Temporal Analysis and Projection of Extreme Particulate Matter (PM10 and PM2.5) Levels Using Association Rules: A Case Study of the Jing-Jin-Ji Region. J. Atmospheric Environment 120,339–50.

Deng,Z. H., 2016. DiffNodesets: An efficient structure for fast mining frequent itemsets. J. Applied Soft Computing. 41,214-223.

Gu,J., Wu,J., Xu,X., et al., 2018. Optimization and Implementation of Parallel FP-Growth Algorithm Based on Spark. J. Computer Applications. 38(11),23-28.

He,L., Ren, Y., 2020. Influence of meteorological factors on power grid load forecasting at different time scales. J. Hydropower and Energy Science.38(09),206-210.

Xu,L., Xu,Y., Wen,J., et al., 2019. Analysis of the relationship between urban and rural residents' electricity load and meteorological factors based on Apriori algorithm. J. Inner Mongolia Meteorology.(2).

Deng,Z. H., 2016. DiffNodesets: An efficient structure for fast mining frequent itemsets. J. Applied Soft Computing. 41,214-223.

Bui,H.,Vo, B., Nguyen,H. et al., 2018.A weighted N-list-based method for mining frequent weighted items. J. Expert Systems with Application. 96, 388-405.

Sadat,Y.K. , Nikaein,T. , Karimipour,F.,2015. Fuzzy spatial association rule mining to analyze the effect of environmental variables on the risk of allergic asthma prevalence. J. Geodesy and Cartography. 41(2), 101-112.

Yin, X., X. Wang, and A. Liu, 2019. *Gas pipeline leak detection based on fuzzy C-means clustering algorithm.* International Journal of Applied Mathematics in Control Engineering. **2**(2): p. 176-181.

Liu,L., Zhang,X., Niu,X., et al., 2019. Research Review of Parallel Association Rules Mining Algorithm Based on Spark. J. Computer Engineering and Applications. 55(09),1-9.

Qingyuan Wang is currently an engineer in Hebei Xiongan Rongwu Expressway Co., Ltd., Baoding, China. He obtained his BS degree from Southwest Jiaotong University, China in 2016. His main research interests are in the areas of Big Data, Data Mining, Cloud Computing and Mechatronics Engineering.

Long Zhang is currently an engineer in Hebei Provincial Commnications Planning, Design and Research Institute Co., Ltd, Shijiazhuang, China. He obtained his MS degree from Shijiazhuang Tiedao University, China in 2014. His main research interests are in the areas of Intelligent Transportation, Cloud Computing.

Xiaoxuan Song is currently an engineer in Hebei Xiongan Rongwu Expressway Co., Ltd., Baoding, China. He obtained his BS degree from Cangzhou Jiaotong College, China in 2016. His main research interests are in the areas of Big Data, Data Mining, Cloud Computing and Communication Engineering.

Honggang Li is currently an engineer in Hebei Xiongan Rongwu Expressway Co., Ltd., Baoding, China. He obtained his BS degree from North China Institute of Science and Technology, China in 2016. His main research interests are in the areas of Big Data, Data Mining, Cloud Computing and Automatic Control.

Wenfeng Qin is currently an engineer in Hebei Xiongan Rongwu Expressway Co., Ltd., Baoding, China. He obtained his BS degree from Jiangxi Vocational and Technical College of Communications, China in 2016. His main research interests are in the areas of Big Data, Data Mining, Cloud Computing and Communication Engineering.

Qingjie Zhao is currently a senior engineer in Hebei Provincial Commnications Planning, Design and Research Institute Co., Ltd, Shijiazhuang, China. He obtained his MS degree from Harbin Institute of Technology, China in 2014. His main research interests are in the areas of Intelligent Transportation, Cloud Computing.

Bingxin Niu is an Assistant Professor in Hebei University of Technology, Tianjin China. He obtained his Ph.D. in computer application from Dalian University of Technology, China in 2019. His research interests include Intelligent Transportation, Mobile Edge Computing, Internet of Things, and Cloud Computing.