文章编号:2096-1472(2023)09-0008-04
DOI:10.19644/j.cnki.issn2096-1472.2023.009.002

基于 Apriori 算法的大气污染物关联性分析研究

郭艳萍,高 云,景 雯

(山西大同大学计算机与网络工程学院,山西 大同 037009)
✉ 38922343@qq.com; 63378161@qq.com; 29708916@qq.com

摘 要:常用的空气质量等级分析方法由于没有考虑大气污染物之间的关联性,导致在治理空气质量时可能存在单一性和片面性。文章提出了基于 Apriori 算法(关联规则算法)对各项大气污染物之间的关联性进行分析研究。该方法对获取的原始样本集进行了属性规约、数据离散化等预处理,将处理后的样本数据集输入模型,设置并调整了模型的最小支持度和最小置信度,直至输出符合现实意义的关联规则集合。根据实验得出的关联规则,证明空气污染问题通常是多种污染物共同作用的结果。

关键词:Apriori;大气污染物;支持度;置信度
中图分类号:TP181 **文献标志码**:A

Research on Correlation Analysis of Air Pollutants Based on Apriori Algorithm

GUO Yanping, GAO Yun, JING Wen

(*School of Computer and Network Engineering, Shanxi Datong University, Datong* 037009, *China*)
✉ 38922343@qq.com; 63378161@qq.com; 29708916@qq.com

Abstract: Commonly used air quality grade analysis methods do not take into account the correlation between atmospheric pollutants, resulting in a potentially one-dimensional and one-sided approach to air quality control. The paper proposes to analyze and study the correlation analysis between various air pollutants based on Apriori algorithm (association rules algorithm). With the proposed method, the obtained original sample set is preprocessed by attribute specification and data discretization, and then, the processed sample data set is input into the model. The minimum support degree and minimum confidence degree of the model are set and adjusted until the output conforms to the association rule set of practical significance. The association rules obtained from experiments prove that air pollution problems are usually the result of a combination of pollutants.

Key words: Apriori; air pollutants; support degree; confidence degree

0 引言(Introduction)

在我国工业快速发展的背景下,随之而来的环境污染问题日益严重。根据产生环境污染原因的不同,引起的污染问题也不同,可分为水源污染、土壤污染、大气污染等。因此,在对待污染问题时不能一概而论,要针对不同类型的污染采用不同的治理措施。此外,由于各地出现的环境污染问题的污染物特点不尽相同,所以通过对污染物进行分类和按照污染物不同的特点,采取不同的措施治理环境污染问题已经成为当今环保工作者重点研究的问题。

山西省大同市属于能源产出城市,大气污染是引起该市环境污染的主要原因之一[1]。目前,我国治理空气污染的主要指标依据是空气质量指数(Air Quality Index,AQI),AQI 值为空气质量分指数(Individual Air Quality Index,IAQI)的最大值,IAQI 值最大的污染物即为首要大气污染物[1-4]。

收稿日期:2023-02-15
基金项目:山西省软科学研究计划(2019041023-5)

Apriori 算法通过对频繁项集进行挖掘,在大数据集上实现了提取关联规则,其主要思想为通过连接的方式生成候选项,计算其支持度,根据支持度进行剪枝,实现频繁项集的生成。空气质量虽然是由 IAQI 值最高的污染物决定其等级,但是多数情况下并不是由单一的污染物作用的。本文使用 Apriori 算法分析引起空气质量变化的多种大气污染物之间的关联性,为针对性地治理大气污染提出了新的思路。

1 Apriori 算法(Apriori algorithm)

Apriori 算法主要是找出事务集中存在的最大频繁 k-项集,并获得最大频繁 k-项集,将其与最小置信度比较后生成强关联规则,即所求关联关系[5]。

1.1 支持度与置信度

关联规则的相对支持度的公式:

$$Support(A\Rightarrow B)=P(A\cap B) \tag{1}$$

即,事务 A 和事务 B 同时发生在事务集中的概率。

置信度公式:

$$Confidence(A\Rightarrow B)=P(A|B) \tag{2}$$

其中,条件概率 $P(A|B)=\dfrac{P(AB)}{P(A)}$,即,如果事务 A 发生,则一定发生事务 B 的概率。

放在事务集中进行研究,事务集中包含事务 A 的个数为事务 A 的支持度计数,也称为事务的计数或频率,从支持度计数推出规则 $A\Rightarrow B$ 的支持度公式:

$$Support(A\Rightarrow B)=\frac{A,B\text{ 同时发生的事务个数}}{\text{所有事务个数}}=\frac{Support\_count(A\cap B)}{Total\_count(A)} \tag{3}$$

从支持度计数推出置信度公式:

$$Confidence(A\Rightarrow B)=P(A|B)=\frac{Support(A\cap B)}{Support(A)}=\frac{Support\_count(A\cap B)}{Support\_count(A)} \tag{4}$$

1.2 关联规则

Apriori 算法分两个步骤实现。

1.2.1 步骤一:寻找最大频繁 k-项集

(1)对所有事务进行扫描,扫描得到的每一项组成候选 1-项集 C_1,并计算每项成员的支持度。

(2)C_1 中各项集的支持度与最小支持度进行比较,将小于等于该阈值的项集剔除后得到频繁 1-项集,记为 L_1。

(3)L_1 与 L_1 连接得到候选 2-项集 C_2,进行剪枝,保留 C_2 中满足约束条件的项集得到频繁 2-项集,记为 L_2。

(4)L_2 与 L_1 连接得到候选 3-项集 C_3,并计算每一项的支持度,进行剪枝,保留 C_2 中满足约束条件的项集得到频繁 3-项集,记为 L_3。

(5)循环以上步骤,得到频繁 k-项集 L_k。

1.2.2 步骤二:由频繁集产生关联规则

步骤一中已经剔除了最小支持度小于等于预设阈值的项集,如果剩下的最小置信度满足预设阈值的规则,那么这些规则就是挖掘到的强关联规则。

2 大气污染物关联模型构建过程(Construction process of air pollutants correlation model)

大气污染物关联性分析模型构建过程如图 1 所示。

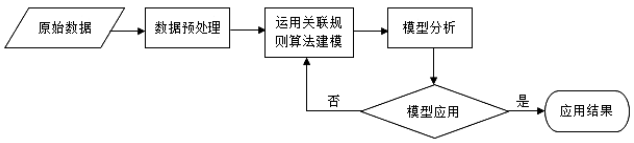

图 1　大气污染物关联性分析模型构建过程

Fig. 1 Construction process of air pollutants correlation analysis model

大气污染物关联性分析主要包括以下步骤。

(1)从相关站点获得大气污染物浓度日报表数据,并将其整理成原始数据。

(2)对大气污染物浓度数据集进行数据预处理,包括数据清洗、属性规约和数据变换等操作。

(3)经过"步骤(2)"形成建模数据,采用 Apriori 算法,设置模型输入参数,获取各大气污染物与空气质量等级之间的关系。

(4)结合实际空气质量划分结果,对模型关联规则结果进行分析,并且将模型挖掘结果应用到实际大气污染物研究中,最后输出获得的关联规则结果。

3 实验过程(Experimentation)

3.1 实验数据准备

本次实验使用的数据集为山西省大同市 2017 年、2018 年和 2021 年三年的空气中各污染物浓度日均值报表数据,共计 1 095 条记录,每条记录包括 $PM_{2.5}$、PM_{10}、NO_2、SO_2、CO、O_3-8h 共 6 种污染物的浓度及其对应的 AQI 和空气质量等级,部分原始数据集及格式如表 1 所示。

表 1　部分原始数据集及格式表

Tab.1 Part of the original data set and format table

日期	*AQI*	质量等级	$PM_{2.5}$	PM_{10}	NO_2	SO_2	CO	O_3-8h
2021-01-01	111	轻度污染	104	111.0	78.75	67.0	100.0	8
2021-01-02	120	轻度污染	119	120.0	87.50	64.0	100.0	9
2021-01-03	116	轻度污染	113	116.0	87.50	66.5	90.0	9
2021-01-04	92	良	92	90.0	83.75	51.0	70.0	23
2021-01-05	114	轻度污染	114	111.0	83.75	52.5	90.0	10
2021-01-06	62	良	62	59.0	45.00	27.0	35.0	37
2021-01-07	34	优	28	34.0	33.75	21.0	30.0	31
2021-01-08	95	良	95	85.0	72.50	57.0	82.5	19
2021-01-09	122	轻度污染	122	97.5	86.25	58.0	82.5	17
2021-01-10	100	良	100	81.5	72.50	50.5	55.0	21
—	—	—	—	—	—	—	—	—

注:$PM_{2.5}$、PM_{10}、NO_2、SO_2、O_3-8h 的浓度单位为 $\mu g/m^3$,CO 的浓度单位为 mg/m^3。

3.2 数据预处理

本实验中数据预处理过程包括数据清洗、属性规约和数据变换。数据来源于站点数据(在观测站点实测到的数据),针对原始大气污染物浓度数据集,经过数据预处理,形成建模数据集。

3.2.1 数据清洗

在站点收集的数据中，存在无效的数据，即数据集中存在某一项或某几项大气污染物浓度为 0 的记录，如表 2 所示。

表2 无效的数据示例表

Tab.2 Invalid data sample table

日期	*AQI*	质量等级	$PM_{2.5}$	PM_{10}	NO_2	SO_2	CO	O_3-8h
—	—	—	—	—	—	—	—	—
2018-03-16	165	中度污染	0	0	51	140	3.4	56
2018-03-17	64	良	0	0	39	86	1.8	79
2018-03-18	92	良	68	113	48	132	3.5	47
—	—	—	—	—	—	—	—	—
2018-03-26	122	轻度污染	0	0	29	60	1.5	97
2018-03-27	70	良	0	0	35	98	1.5	61
2018-03-28	496	严重污染	0	0	48	236	3.2	73
—	—	—	—	—	—	—	—	—

注：$PM_{2.5}$、PM_{10}、NO_2、SO_2、O_3-8h 的浓度单位为 $\mu g/m^3$，CO 的浓度单位为 mg/m^3。

大同市的实际空气质量情况是大气污染物浓度长期可能较低，但基本不存在污染物浓度都为 0 的情况，为了提高模型分析的准确性，需要对其进行处理，在原始数据集中将大气污染物浓度为 0 的记录直接删除，获得有效数据集。本次实验的原始数据集包含 1 095 条记录，删除 31 条无效数据后，有效数据集包含 1 064 条记录，数据有效率约为 97%。由此可见，原始数据集的数据可靠性较高。

3.2.2 属性规约

从表 1 可知，原始样本集数据共有 9 个属性，为了更有效地分析大气污染物之间的关联性，将其中与实验任务无关的属性剔除。经过分析可得，属性“日期”与“*AQI*”与本次关联分析无关，因此选取其余 7 个属性值构成数据集进行分析。属性规约后的数据集如表 3 所示。

表3 属性规约后的部分数据

Tab.3 Partial data after attribute specification

质量等级	$PM_{2.5}$	PM_{10}	NO_2	SO_2	CO	O_3-8h
Ⅲ	78	172	63	84	4.0	16
Ⅲ	90	190	70	78	4.0	17
Ⅲ	85	182	70	83	3.6	17
Ⅱ	68	130	67	52	2.8	46
Ⅲ	86	172	67	55	3.6	20
Ⅱ	44	68	36	27	1.4	73
Ⅰ	19	34	27	21	1.2	61
Ⅱ	71	120	58	64	3.3	37
Ⅲ	92	145	69	66	3.3	34
Ⅱ	70	127	45	58	3.1	38
—	—	—	—	—	—	—

注：$PM_{2.5}$、PM_{10}、NO_2、SO_2、O_3-8h 的浓度单位为 $\mu g/m^3$，CO 的浓度单位为 mg/m^3。

3.2.3 数据变换

本实验主要采用属性构造和数据离散化两种方法进行数据变换。首先进行属性构造，获得各项大气污染物的 IAQI 值，然后离散化处理数据集，得到建模数据，该操作使用聚类算法完成。

(1)属性构造。原始样本集中的各种大气污染物的属性值描述的是污染物浓度，但是每种污染物浓度的量纲不同，所以只看污染物浓度值是没有意义的，空气质量等级依赖每种污染物的 IAQI 值，因此需要将污染物浓度转换为其对应的 IAQI 值。计算污染物项目 P 的 IAQI 值公式如下：

$$IAQI_P=\frac{IAQI_{HI}-IAQI_{LO}}{BP_{HI}-BP_{LO}}(C_P-BP_{LO})+IAQI_{LO} \quad (5)$$

其中，$IAQI_P$ 为污染物项目 P 的 IAQI 值；C_P 为污染物项目 P 的浓度值；BP_{HI} 与 BP_{LO} 分别为污染物项目 P 与 C_P 相近的污染物浓度高位限值与低位限值；$IAQI_{HI}$ 与 $IAQI_{LO}$ 分别为污染物项目 P 与 BP_{HI}、BP_{LO} 对应的高位值与低位值。

针对表 1 中各污染物浓度进行属性构造转化为 IAQI 值后的数据集如表 4 所示。

表4 属性构造后的数据集

Tab.4 Data set after attribute construction

质量等级	$PM_{2.5}$	PM_{10}	NO_2	SO_2	CO	O_3-8h
Ⅲ	58	132	58	77	3.5	11
Ⅲ	70	150	65	71	3.5	12
Ⅲ	65	142	65	76	3.1	12
Ⅱ	48	90	62	45	2.3	41
Ⅲ	66	132	62	48	3.1	15
Ⅱ	24	28	31	20	0.9	68
Ⅰ	10	16	22	14	0.7	56
Ⅱ	51	80	55	57	2.8	32
Ⅲ	72	105	64	59	2.8	29
Ⅱ	58	132	58	77	3.5	11
—	—	—	—	—	—	—

(2)数据离散化。由于 Apriori 算法只适用于离散数据，无法对连续数值进行处理，即处理数据为 A、B、C 的类别值，而不是数字，因此为了将 IAQI 数据值转换为适合 Apriori 建模的格式，需要将数据进行离散化。本实验采用聚类算法对各污染物的 IAQI 值进行离散化处理，空气质量根据受污染程度分为 6 个等级，因此将每个属性聚成 6 类，并将每两类中心的均值作为其边界点。聚类中心第一列的值设为 0，使用边界值及聚类的个数得到结果如表 5 所示。

表5 边界值及聚类个数表

Tab.5 Table of boundary values and number of clusters

类别	等级					
	1	2	3	4	5	6
A	0	27.577 48	46.917 68	70.514 15	101.565 40	163.141 50
An	300	367	255	96	41	5
B	0	31.093 26	48.288 04	66.738 99	90.948 54	262.137 90
Bn	122	228	391	264	58	1
C	0	20.511 49	29.600 04	39.303 16	50.539 83	65.875 80
Cn	177	246	268	197	137	39
D	0	18.416 68	29.622 37	43.236 27	57.695 45	76.332 39
Dn	363	283	198	123	74	23
E	0	21.085 22	31.363 94	44.230 08	60.826 13	79.875 67
En	308	292	229	143	53	39
F	0	25.120 25	40.440 03	56.893 31	78.726 93	105.912 70
Fn	137	311	248	200	122	46

将 6 类边界值求出后，即可将 IAQI 属性值离散成为 6 种类别号，离散后的部分结果如表 6 所示。

表6 离散化后数据表

Tab.6 Data sheet after discretization

质量等级	$PM_{2.5}$	PM_{10}	NO_2	SO_2	CO	O_3-8h
Ⅲ	A5	B5	C6	D5	E6	F1
Ⅲ	A5	B5	C6	D5	E6	F1
Ⅲ	A5	B5	C6	D5	E6	F1

续表

质量等级	$PM_{2.5}$	PM_{10}	NO_2	SO_2	CO	O_3-8h
Ⅱ	A4	B4	C6	D4	E5	F1
Ⅲ	A5	B5	C6	D4	E6	F1
Ⅱ	A3	B3	C4	D2	E3	F2
Ⅰ	A2	B2	C3	D2	E2	F2
Ⅱ	A4	B4	C6	D4	E6	F1
Ⅲ	A5	B5	C6	D5	E6	F1
Ⅱ	A4	B4	C6	D4	E4	F1
—	—	—	—	—	—	—

3.3　模型构建

本实验的目标是探索大气污染物浓度之间以及各污染物与空气质量等级之间的关联关系,因此采用 Apriori 算法寻找上述关联关系。

关联规则模型的主要用途是在数据集中挖掘其中数据项相互之间是否具有关联关系。Apriori 算法基于数据项的统计规律,寻找数据项之间隐藏着的未知关系,并分析其关联规则。根据分析得到的关联性,可以根据已知属性的信息对其他属性的信息进行推断[6]。当置信度满足阈值需求时,则认定为该规则成立。

3.3.1　大气污染物关联规则模型

大气污染物关联规则模型流程图如图 2 所示。

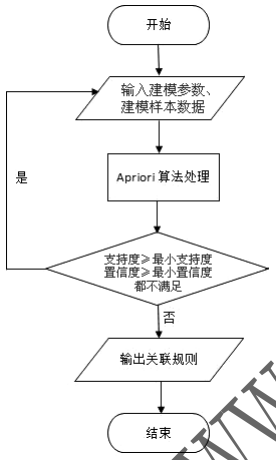

图 2　大气污染物关联规则模型

Fig. 2 Air pollutants association rule model

从图 2 可以看出,模型主要由输入、算法处理、输出部分组成;输入部分包括建模参数的输入和建模样本数据的输入;算法处理采用 Apriori 算法;输出结果为关联规则的结果。

模型建立具体步骤如下[7-8]。

(1)设置建模初始最小支持度和初始最小置信度。

(2)模型样本数据集的输入。

(3)以设置的模型参数和现实分析目标作为条件,使用 Apriori 算法分析建模数据集,根据得到关联规则执行“步骤(4)”或“步骤(5)”。

(4)如果所有挖掘到的规则都不满足条件,则对模型参数进行调整,返回“步骤(3)”执行。

(5)挖掘到的规则满足条件且符合现实意义,输出关联规则结果集。

实际应用中,对于最小支持度与最小置信度的设置并不是固定值,其中大部分应用都是根据实际需求和历史经验设置初始值,然后经过多次调整,获取与实际应用情况相符的关联规则结果。本实验设置模型的初始输入参数为最小支持度=0.05、最小置信度=0.6,得出关联规则数目较多且规则不明显。经过多次调整并结合实际业务分析,最终获得的实验结果是在最小支持度=0.06、最小置信度=0.75 的条件下得出的。

3.3.2　模型结果分析

根据上述模型运行结果,得出了 89 条关联规则,去除 48 条无意义的规则,有效规则为 41 条,部分有效规则如表 7 所示。

表7　模型运行部分结果表

Tab.7 Partial results tale of the model operation

有效规则	支持度	置信度
A2-F4-Ⅱ	0.079 452 055	1
B3-F4-Ⅱ	0.069 406 393	1
B2-F4-Ⅱ	0.063 926 941	1
A3-C5-Ⅱ	0.064 840 183	0.986 111 111
C2-F4-Ⅱ	0.063 013 699	0.985 714 286
E2-F4-Ⅱ	0.071 232 877	0.975 000 000
A1-F4-Ⅱ	0.065 753 425	0.972 972 973
E1-F4-Ⅱ	0.064 840 183	0.972 602 740
D1-F4-Ⅱ	0.091 324 201	0.961 538 462
B2-F2-Ⅰ	0.062 100 457	0.957 746 479
—	—	—

分析模型运行结果总结如下。

(1)大同市空气质量近几年较好,空气质量等级分布如图 3 所示。

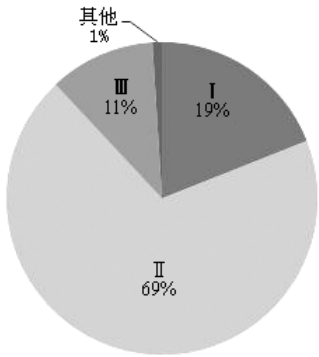

图 3　空气质量等级分布图

Fig. 3 Distribution chart of air quality level

从图 3 可以看出,数据集中(1 095 条数据)记录的空气质量等级基本集中在Ⅰ级和Ⅱ级,Ⅲ级占 11%,Ⅲ级以下占比为 1%。因此,使用 Apriori 算法获得的规则也主要为Ⅰ级和Ⅱ级的规则,其中在 41 个有效规则中,39 个为Ⅱ级,2 个为Ⅰ级,规则分布符合实际情况。

(2)以表 7 第一个规则为例,A2-F4-Ⅱ的支持度约为 0.079 5,置信度为 1,说明 $PM_{2.5}$ 的 IAQI 值处于(25.577 48,46.917 68]范围内,O_3-8h 的 IAQI 值处于(56.893 31,78.726 93]范围内,空气质量等级为Ⅱ级的可能性为 100%,而发生这种情况的可能性约为 7.95%。分析 41 个有效规则可知,大气污染物的占比分布如图 4 所示。

(下转第 32 页)