doi:10.3772/j.issn.2095-915x.2017.02.006

企业大气污染排放数据挖掘方法研究
——以北京及其周边为例

北京石油化工学院信息管理与信息系统系 北京 102617
彭珍 卜潇箫 高雅 赵宁 明冲

摘 要 目前，大气污染决策已经成为我国一项直接影响经济、社会发展的重要情报工作。为了得到对北京及其周边主要工业企业大气污染排放的精确、可信的情报，本研究通过对北京及其周边企业大气污染排放数据的获取，构建企业大气污染排放数据库，并对北京及其周边地区不同类型企业、不同类型污染物以及不同方位的污染排放进行数据分析，寻找造成北京及其周边大气污染的主要污染物、主要污染类型企业和主要方位，再基于模糊认知图的数据挖掘方法定量地建立企业污染排放对大气污染的非线性影响关系，由此为北京及其周边企业大气污染防控提供决策支持。

关键词： 大气污染，企业大气污染排放数据，模糊认知图，数据挖掘
中图分类号： G35

Research on Data Mining Method for Enterprise Air Pollution Emission —A Case Study of Beijing and Its Surrounding Regions

Department of Information Management and Information System, Beijing Institute of Petrochemical Technology, Beijing 102617, China
PENG Zhen BU XiaoXiao GAO Ya ZHAO Ning MING Chong

Abstract At present, air pollution decision has become one of the most important information services,

基金项目： 本文受国家自然科学基金项目“探索雾霾形成的模糊认知图构建及其数据挖掘方法”(71601022)；北京市社科基金项目“数据驱动的北京大气污染源智能分析”（15JDJGB028）；北京市自然科学基金项目（4173074）、北京市URT项目（2016J00152、2016J00153）的资助。

作者简介： 彭珍（1981–），博士，副教授，研究方向：模糊认知图、数据挖掘；卜潇箫（1996–），学士，研究方向：数据分析；高雅（1996–），学士，研究方向：数据分析；赵宁（1997–），学士，研究方向：数据分析；明冲（1997–），学士，研究方向：数据分析。

doi:10.3772/j.issn.2095-915x.2017.02.006

which directly affects the economic and social development in our country. In order to obtain accurate and credible information of atmospheric pollution emissions from the major industrial enterprises in Beijing and its surrounding areas, a kind of enterprise air pollution emission database was constructed through acquisition of air pollution emission data from Beijing and its surrounding enterprises. And data analysis was carried out on polluting emission from different types of enterprises, different types of pollutants and different directions for detecting the major pollutants, the major pollution types of enterprises and major directions resulting in air pollution of Beijing and its surrounding regions. On the basis of these analysis, the nonlinear relationships between the major pollution enterprises and air pollution were mined quantitatively based on fuzzy cognitive map, which can provide decision support for air pollution emission control from enterprises in Beijing and its surrounding regions.

Keywords：Air pollution, enterprise air pollution emission data, fuzzy cognitive map, data mining

1 引言

情报是由情报的获得与传播机构向决策者提供的，决策者利用获得的情报选择最佳的行动路径[1]。随着现代社会信息化程度的迅猛发展，情报工作面临着两大方面的问题：一方面是各行业数据产生量的剧增；另一方面是决策者需求情报的专业化、精确化。因此，如何处理日益增长的数据信息，为决策者提供更精确的情报已迫在眉睫。这就要求情报研究与服务工作必须基于情报信息技术的支持，利用数据采集、分析、挖掘等方法与手段，以提高决策信息的精度和可信度、满足情报用户的需求。

数据挖掘就是从大量数据中抽取、分析和模型化处理，从中发现隐含在数据中概念、模式、规律等有用的知识[2]，得到辅助决策的关键性信息。模糊认知图[3,4]是1986年在认知图的基础上，把概念间具有的三值逻辑关系扩展为区间[-1,1]的模糊关系上。它作为一种知识表示和智能建模方法，充分利用了模糊逻辑的模糊信息处理能力、认知图的因果关系的传播方法和神经网络的动态自适应特性，能够很好地将三者结合在一起，基于模糊认知图的数据挖掘方法有利于数据的智能化仿真与处理。

大气污染决策就是一项非常重要的情报工作。特别是随着经济的不断增长，环境污染越来越严重，大气污染问题直接对我国社会事业的和谐构成了非常严重的威胁和挑战，且已经成为影响经济、制约社会的重大问题。研究表明，仅北京一个城市，每年因环境污染造成的损失就高达116亿多元，而这并未包括自然资源的损失，也没有计入生态破坏的损失。尤其是大气污染，对北京市造成的经济损失约95.2亿元，占总污染造成损失的81.75%[5]。

根据调查研究，大气污染问题不仅是一个行政区的问题，还是一个区域性的问题，北京周边区域的大气污染对北京地区的污染有着不同程度的影响[6]。区域大气污染排放的情报决策已成为区域大气污染防控的一个关键的手段。

目前，大气污染情报分析[7-10]主要有两种方式。一种是“至下而上”的源解析方法，这种方法从污染源出发，基于污染源排放清单，

通过对污染源的分析来识别污染源。污染源排放清单的建设费时费力，建设的最低层是行业 / 产业污染源，通过大气污染排放清单是无从获得更底层企业级污染排放的情况。另一种是“至上而下”的源解析方法，该方法是从污染物出发，主要是通过大气颗粒物化学成分和物理特性来推断污染物来源，估算各类污染源的贡献率。这类方法采用质量守恒的原理，得到各类燃料级污染源的排放情况，仍是无从得到企业级的污染排放。

而且，大气污染排放情报分析也主要有宏观和微观两个层面进行。一种宏观的研究。很多研究采用数值模拟方法对仪器的监测数据，一般是对颗粒物（PM10、PM2.5）的浓度进行研究[11-13]，得到北京污染来自区域性输送和周边局地城市污染的双重贡献。这些研究是对北京及其周边的宏观研究，是从污染排放数值的模拟仿真研究周边对北京地区的影响，缺少对污染源企业的角度进行大气污染的分析与发现。另一种是微观的研究。目前，企业污染排放的研究主要集中在不同类型企业排放的微观研究上[14-16]，有定性的分析火电厂企业排放、钢铁企业排放的污染物有哪些，也有定量的估算（基于用电量）工业企业直接污染排放量是多少。缺少的是对区域内不同类型企业、不同类型污染物污染排放的中观分析。

针对这些问题，本文拟从企业污染源的角度，采用数据挖掘的思路，从源头上对企业大气污染排放信息进行获取、分析和基于模糊认知图的数据挖掘方法，为明确北京大气污染来源、加强北京大气污染防治提供决策支撑。

2 模糊认知图

模糊认知图（Fuzzy Cognitive Map，FCM）是一种具有语义的图模型，具有直观的知识表达能力、强大的基于数字矩阵的推理机制等优点。FCM 把知识蕴含在概念结点及概念结点间的关系中，通过概念间的关系来模拟模糊推理，通过整个图中各概念结点的相互作用来模拟系统的动态行为。FCM 与神经网络模型相比，模糊认知图模型的优势在于：

（1）模糊认知图模型中每一个结点和弧都有很强的语义，从而使整个图结构呈现很强的语义，推理的结果易于解释。神经网络是一种数值框架，不能直接表示结构知识，无法说明预测机理。

（2）模糊认知图模型可以应用专家知识在一定程度上来弥补学习数据的不足，其推理是基于矩阵运算，所以符合智能行为由数据驱动的人工智能（Artificial Intelligence，AI）发展方向。神经网络没有利用系统本身的专家知识，并需要大量的训练数据，对于不容易得到训练数据的系统，神经网络作用十分有限。

（3）模糊认知图模型不仅可以表示语义网络，而且还可以处理分布知识，对于任意数量的知识源可以分别构造自己的认知图，并可以进行相互任意的叠加，从而得到整个知识源的一个联合知识分布。神经网络不具有叠加性，其对于大型系统的学习能力较差。

（4）模糊认知图模型不仅可以用来预测，还可以用于因素变动的敏感性分析、解释因果关系、策略规划等。神经网络属于黑箱模型，无法实现对整个系统的结构化分析。

doi:10.3772/j.issn.2095-915x.2017.02.006

FCM 状态空间最初由初始条件决定，经阈值 / 演变函数推理，通过整个网络各概念结点间的相互作用模拟交互网络的动态行为，它自身相当于一个非线性动力系统。模糊认知图状态终止于固定点 (Fixed Point)、极限环 (Limited Circle) 或 “ 混沌 ” 的吸引子，系统的规则存储在空间自身。

基于 FCM 的数据挖掘方法能够从历史数据中学习发现 FCM 中概念之间的关联权重，逐步发展为构造 FCM 的主流趋势。FCM 挖掘算法 [17-20] 主要包括非线性 Hebbian 算法 (Nonlinear Habbian Learning, NHL)、遗传算法 (Genetic Algorithm,GA) 等。由于 FCM 直观的知识表达、自动从数据中学习的能力以及它与神经网络、图论、模糊逻辑等领域的密切联系，使得 FCM 已成功地应用到工程、医学、管理等各个领域 [20-23]，比如系统评估、故障诊断、医疗决策、旅游行为分析等。

3 企业大气污染排放数据获取与分析

3.1 北京及其周边企业大气污染排放数据获取

本研究以北京及其周边为例，北京位于华北平原西北边缘，地处东经 113°27′ ~ 119°50′，北纬 36°05′ ~ 42°40′ 之间，北部环山，被河北所环抱。自京津冀协同发展规划以来，所在北京很多污染企业都已陆续搬离北京，迁至周边河北。在所调查的这些企业中，共涉及 94 家大气污染排放企业，其中 76 家企业来自河北，只有 18 家企业来自北京，企业排放的大气污染物共计近 20 种。这些工业企业大气污染来自环保局 2015 年工业企业监督性监测数据 [24-25]。

为了便于对企业污染排放数据的有效管理，本研究对企业类型进行分类。一是对企业所在地理方位的分类，根据企业所在地区可以识别；二是对企业类型的分类，根据企业加工、处理主要对象的不同，将所调研企业分为 12 个种类，见表 1 所示。

为了进一步对企业大气污染排放监测信息实施数据分析，在理解和分析企业大气污染排放的情况下，建立了企业大气污染排放数据库，其逻辑模型 E-R 图如图 1 所示。它描绘了企业污染排放系统的中包含哪些实体（企业地区、企业类型、企业、污染物、污染排放标准）、实体之间的静态联系（污染监测）、以及实体的特征属性。在企业大气污染排放数据库中污染监测是核心，它是连接企业、污染物的关键实体，包括有监测位、监测时间、排放浓度等若干属性。

表1 调研企业分类表

企业类型	企业数量	百分比
化工	6	6.38%
钢铁	30	31.91%
有色金属	11	11.70%
煤炭	6	6.38%
电池	7	7.45%
建材	3	3.19%
火电	10	10.64%
水泥	6	6.38%
电子材料	8	8.51%
造纸	1	1.06%
制革	1	1.06%
回收加工	5	5.32%

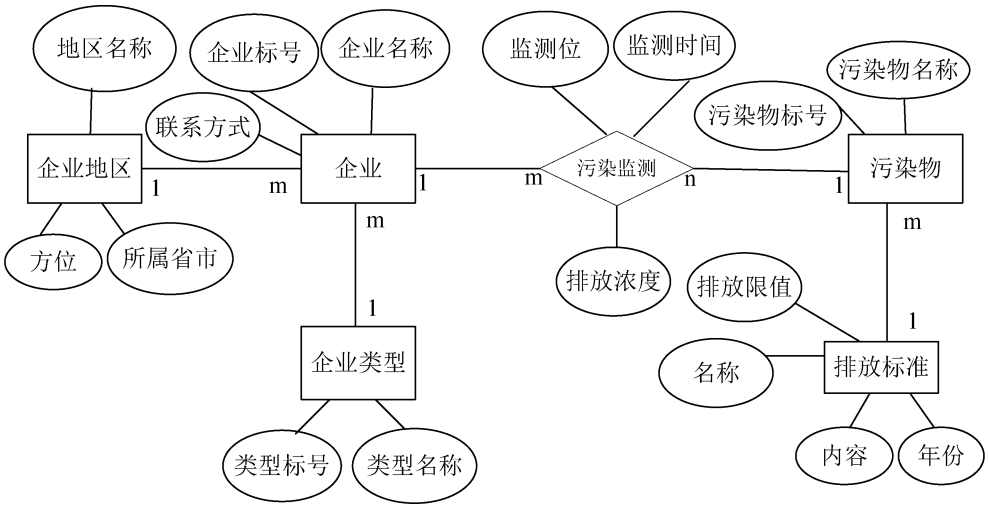

图1 企业大气污染排放数据库E-R图

3.2 北京及其周边企业大气污染排放数据分析

本研究基于数据库的结构化查询语言（Structured Query Language, SQL），分析得到在不同区域方位、不同类型企业、不同污染物的企业污染排放数据，再对比分析从而得到影响北京及其周边大气污染的主要污染物、主要污染企业类型和主要方位。

从北京不同方位、不同类型的企业来观测分析污染排放情况，如图 2 所示，东部的钢铁与煤炭企业，东南部的钢铁与火电企业大气污染排放量较大。

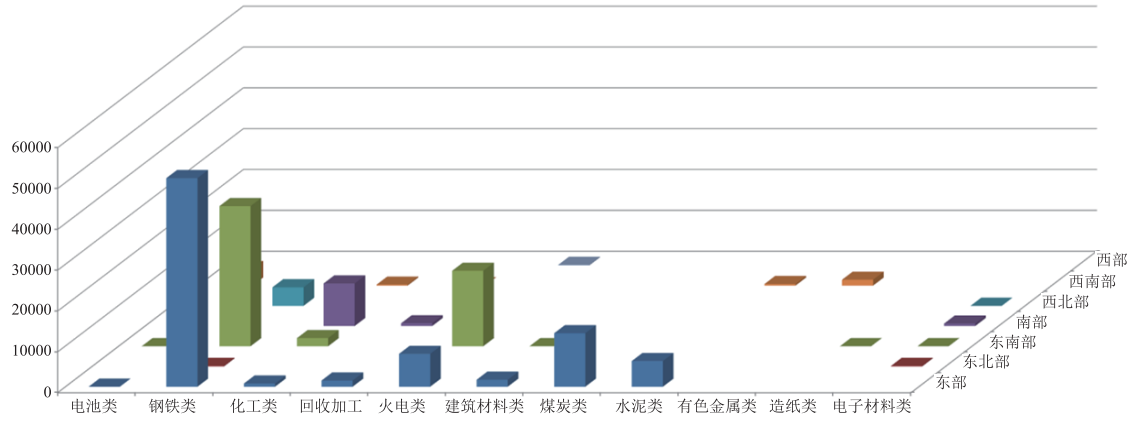

图2 北京及其周边不同方位不同类型企业污染物排放对比

从不同方位看不同类型的污染物排放情况，如图 3 所示，北京东部、东南部区域的 NO_x、SO_2、PM 排放量最为显著。SO_2 与 NO_x 均为有害物质，尤其是 NO_x 毒性大，并且不容易扩散，危害性更大。NO_x、SO_2 又是产生颗粒物（PM）二次污染的主要来源。且 SO_2、NO_x 和

doi:10.3772/j.issn.2095-915x.2017.02.006

颗粒物（PM）这三项正是雾霾的主要组成，这也说明了企业污染排放对雾霾形成的贡献不容小视。在北京地区除了企业污染排放外，汽车也是 NO_x 的重要来源。所以，NO_x 和 SO_2 是北京地区大气污染最主要的危害，也应是大气污染去除的最主要对象。

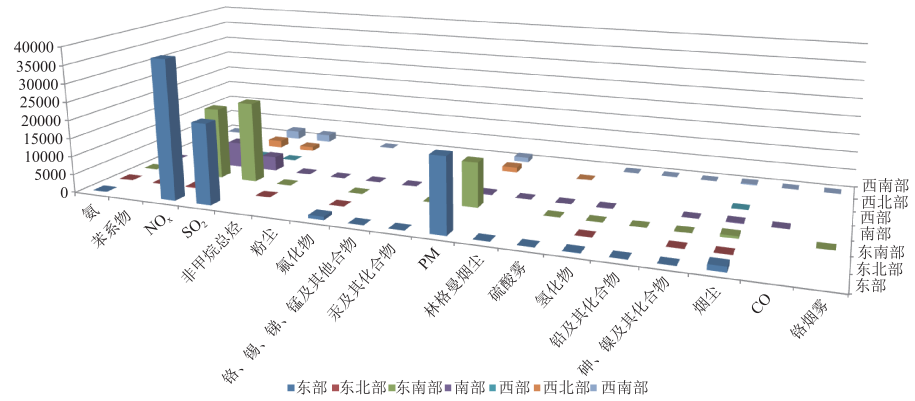

图3 北京及其周边不同方位企业大气污染物排放对比

从北京及其周边不同类型企业不同类型污染物排放量看，如图 4 所示，钢铁企业的污染排放最为突出，特别是钢铁企业的 NO_x、PM 与 SO_2 的排放浓度较大。其次是火电类企业的 SO_2、NO_x 排放量也不容乐观。

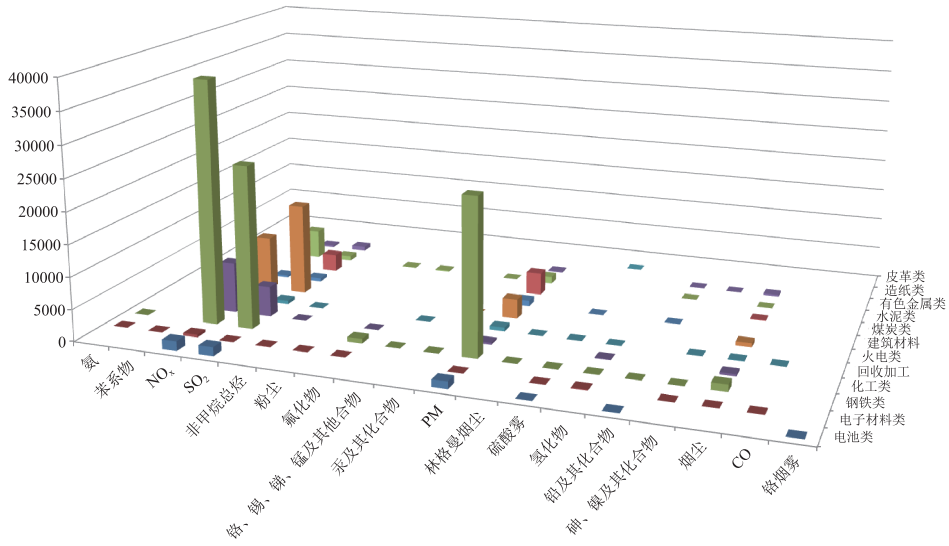

图4 北京及其周边不同类型企业大气污染物的排放对比

值得注意的是，北京所有企业的污染物的排放量都达到了国家的排放标准[26]，而北京周边个别企业中在四种空气污染物排放上超出国家标准标准，它们依次是铅及化合物、NO_x、SO_2、烟尘。有污染物超标排放的企业占 11% 的比例。以铅及化合物为例，它的排放量相比 NO_x、SO_2 等要少很多，但是个别企业的铅及化合物排放浓度远超国家标准值。

4 基于模糊认知图的企业大气污染排放数据挖掘方法

4.1 企业大气污染排放的模糊认知图构建

基于模糊认知图理论，本研究定义企业大气污染排放的模糊认知图模型

$$U=(C, W, A, f) \quad (1)$$

其中：

- $C=\{C_1,C_2,\ldots,C_n\}$ 是模糊认知图中的结点集合，表示为企业大气污染中企业污染概念和区域大气污染概念的集合。
- $W=\{w_{ij} \mid w_{ij}$ 是有向边 $<C_i, C_j>$ 的权值 $\}$（即 w_{ij} 表示结点 C_i 对 C_j 的因果关联强度）。$w_{ij}=w(C_i \rightarrow C_j)$，如果 $w_{ij}>0$，则 C_i 对 C_j 有正影响，即 C_i 的增加（或减少）引起 C_j 的增加（或减少）；如果 $w_{ij}<0$，则 C_i 对 C_j 有负影响，即 C_i 的增加（或减少）引起 C_j 的减少（或增加）；如果 $w_{ij}=0$，则表明 C_i 对 C_j 没有影响，此时 C_i 到 C_j 不联边。这里表示为企业大气污染对区域大气污染的影响权重。
- $A(t)=\{A_1(t), A_2(t), \ldots, A_n(t)\}$ 是结点在 t 时刻的状态序列。A(0) 是所有结点的初始状态，A(t) 是所有在 t 步骤的状态。一个结点在 t 步骤的状态与前一步骤的结点状态有关，受到与之有因果关系的概念影响。在北京及其周边企业大气污染排放系统中，每个概念结点在不同时刻都有一个概念状态，表示为各概念在某时间段的大气污染排放水平集合。
- f 是一个转换函数或阈值函数（Threshold Function），是 A(t+1) 与 A(t) 之间状态转换的函数。通过模拟企业大气污染排放的模糊认知动态关系，可定量地观测污染排放的非线性影响关系。

$$\forall i, j \in \{1,2,\ldots,n\}, \qquad A_i(t+1)=f\left(\sum_{\substack{i=1,\\ i\neq j}}^{n} w_{ji}A_j(t)\right) \quad (2)$$

f 一般选用二值（3）、三值（4）或 S 型曲线函数（5）、（6）：

二值函数

$$f(x)=\begin{cases}0 & x\le 0\\ 1 & x>0\end{cases} \quad (3)$$

三值函数

$$f(x)=\begin{cases}-1 & x\le -0.5\\ 0 & -0.5<x<0.5\\ 1 & x\ge 0.5\end{cases} \quad (4)$$

S型曲线函数I

$$f(x)=\frac{1}{1+e^{-cx}} \quad (5)$$

S型曲线函数II

$$f(x)=\frac{1-e^{-cx}}{1+e^{-cx}} \quad (6)$$

FCM 每个结点都有一个状态空间，FCM 在时间段 t 的状态可以方便地用向量函数表示。其中 $A_i(t)$ 为结果结点即 t 时间段的区域大气污染程度，为企业大气污染的输出，$A_j(t)$ 为企业大气污染概念结点在 t 时间段的污染程度，w_{ji} 为企业大气污染排放对区域大气污染的影响程度。

4.2 基于模糊认知图的企业大气污染排放数据挖掘算法及在北京及其周边的应用

根据对北京及其周边企业大气污染排放的数据分析，本研究选择主要污染企业：钢铁企业、

doi:10.3772/j.issn.2095-915x.2017.02.006

化工企业、火电企业、煤炭企业为数据挖掘对象。基于企业大气污染排放的模糊认知图模型，选用非线性 Hebbian（NHL）模糊认知图学习算法为企业大气污染排放的数据挖掘方法。因为本研究中的输出概念的状态即区域的大气污染程度无法获知，故选用非线性 Hebbian（NHL）模糊认知图学习算法这种无监督的学习机制进行数据挖掘。

基于非线性 Hebbian（NHL）的模糊认知图学习算法的基本思想是，根据系统的初始状态样本数据，使得输出概念结点经激活函数所产生的数学期望最大，由此得到下述学习目标如公式 (7)。

$$\max J(w) = E[A_i^2] \quad (7)$$

式中，w 为的影响程度的集合，即影响权重为待优化的变量。

根据非线性 Hebbian（NHL）学习的原理，影响权重受 ||w||=1 的限制，定义 FCM 调节权值公式为 (8)。

$$\Delta w_{ji}(t+1) = \eta A_i(t)(A_j(t) - w_{ji}(t)A_i(t)) \quad (8)$$

其中，η 是一个很小的正标量因子，称为学习速度参数。

具体的学习算法步骤为：

Step1：设置循环次数和学习参数 η 和 c

Step2：读初始状态 A(0) 和初始权矩阵 w(0)。

Step3：重复每个迭代步骤 t：

1：由式 (2) 计算 A(t)；

2：由式 (8) 更新 $w_{ji}(t)$；

3：由 (7) 判断是否为最大值，保存当前最大值。

Step 3：直至满足循环次数结束。

Step 4：返回收敛区域内的最终的影响权重 w^{update}。

本研究选用 2015 年北京与河北工业污染企业的每月排放数据为训练样本，基于模糊认知图的非线性 Hebbian（NHL）学习算法。使用 matlab 工具，选用公式 (6) 为状态转换函数，设置循环次数为 500000 次，学习参数 η=0.1、c=1，得到北京及其周边企业（钢铁企业、化工企业、火电企业、煤炭企业）大气污染对区域污染的影响权重即 w：

$$w = [0.7749, 0.0001, 0.6321, 0.0001]$$

实验证明，该影响权重是稳定的，它的结果与初始权重无关。该影响权重定量地、精确地反映了各类污染企业对区域大气污染的影响程度。

5 结论

北京作为国家的首都，政治、文化中心，企业大气污染问题目前已经成为一个急需解决的问题之一。本研究从企业污染数据出发，对北京及其周边主要工业企业大气污染进行了数据采集、分析与挖掘，基于模糊认知图的企业大气污染数据挖掘方法实现了企业大气污染的模糊认知图构建及其影响程度的数据挖掘，本研究更明确了北京及周边企业的大气污染情况，为北京及其周边企业大气污染智能分析提供了决策支撑，为提升企业大气污染防控和大气污染排放预警建设提供基础情报支持。据此，在加强北京区域大气污染防治与治理工作中，提出以下建议：

（1）要区域统筹统一治理，特别需要加强

北京周边企业污染的治理工作，如东、东南部尤其要重视，要统筹规划，要联合治理，考虑空气流动、季节因素等等统一协调。

（2）必须全力推进企业，特别是钢铁、火电等类型企业，环保设备的升级改造，严格控制生产作业中的污染排放。

（3）提升大气污染防控标准，特别是氮氧化物（NO_x）、二氧化硫（SO_2）、颗粒物（PM）的排放浓度标准，同时还应加强对企业大气污染排放的预警，可从不同方位区域内企业的污染排放预警、企业自身的污染排放预警、区域/企业不同污染物排放浓度预警三个层面完成企业污染排放的预警和防范。

参考文献

[1] 黄彦敏，孙成权，吴新年 . 国内外科技战略情报研究现状及我国的发展建议 [J]. 图书与情报，2007(1): 86-88.

[2] Jiawei Han, Micheline Kamber 著，范明，孟小峰译 . 数据挖掘概念与技术 [M]. 北京：机械工业出版社，2008.

[3] Kosko B. Fuzzy Cognitive Maps[J]. International Journal of Man-Machine Studies, 1986, 24(1):65-75.

[4] Peng Z, Wu L, Chen Z. Research on Steady States of Fuzzy Cognitive Map and its Application in Three-Rivers Ecosystem[J]. Sustainability, 2016, 8(1):40.

[5] 王祎俊 . 北京市空气污染与经济发展关系研究 [J]. 人口与经济，2010(S1):180-181.

[6] 胡芳芳 . 北京市空气污染的空间统计分析 [D]. 北京：首都经济贸易大学，2010.

[7] 谈佳妮，余琦，马蔚纯，等 . 小尺度精细化大气污染源排放清单的建立 —— 以上海宝山区为例 [J]. 环境科学学报，2014, 34(5):1099-1108.

[8] 孔少飞 . 大气污染源排放颗粒物组成、有害组分风险评价及清单构建研究 [D]. 天津：南开大学，2012.

[9] 张延君，郑玫，蔡靖，等 . PM_(2.5) 源解析方法的比较与评述 [J]. 科学通报，2015(2):109-121.

[10] 邱立民 . 城市大气中颗粒物源解析的不确定性研究 [D]. 长春：吉林大学，2012.

[11] 马锋敏 . 北京及周边地区典型大气污染过程的数值模拟研究 [D]. 南京：南京信息工程大学，2007.

[12] 王占山，李云婷，孙峰，等 . 2015 年 1 月下旬北京市大气污染过程成因分析 [J]. 环境科学学报，2016, 36(7):2324-2331.

[13] 吴莹，吉东生，宋涛，等 . 夏秋季北京及河北三城市的大气污染联合观测研究 [J]. 环境科学，2011, 32(9):2741-2749.

[14] 陈增锋 . 钢铁企业大气污染源数据库系统及污染预测系统的开发 [D]. 武汉：武汉科技大学，2010.

[15] 伯鑫，王刚，温柔，等 . 京津冀地区火电企业的大气污染影响 [J]. 中国环境科学，2015, 35(2):364-373.

[16] 张一民，范金松，赵燕生，等 . 大型化工企业大气污染天气动态预测系统的研究 [J]. 气象科学，1994(4):369-375.

[17] Papakostas G A, Koulouriotis D E, Polydoros A S, et al. Towards Hebbian Learning of Fuzzy Cognitive Maps in Pattern Classification Problems[J]. Expert Systems with Applications, 2012, 39(12):10620-10629.

[18] Stach W, Kurgan L, Pedrycz W. A Divide and Conquer Method for Learning Large Fuzzy Cognitive Maps[J]. Fuzzy Sets & Systems, 2010, 161(19):2515-2532.

[19] Peng Z, Wu L, Chen Z. NHL and RCGA Based Multi-Relational Fuzzy Cognitive Map Modeling for Complex Systems[J]. Applied Sciences, 2015, 5(4):1399-1411.

[20] Peng Z, Peng J, Zhao W, et al. Research on FCM and NHL Based High Order Mining Driven by Big Data[J]. Mathematical Problems in Engineering, 2015, 2015:1-7.

doi:10.3772/j.issn.2095-915x.2017.02.006

[21] Lee K S, Kim S H. Fault Diagnostic System based on Fuzzy Time Cognitive Map[J]. Transaction on Control Automation & Systems Engineering, 1999, 1.

[22] Papageorgiou E I. A new methodology for Decisions in Medical Informatics using Fuzzy Cognitive Maps based on Fuzzy Rule-extraction Techniques[J]. Applied Soft Computing, 2011, 11(1): 500-513.

[23] León M, Mkrtchyan L, Depaire B, et al. Learning and Clustering of Fuzzy Cognitive Maps for Travel Behavior Analysis[J]. Knowledge and Information Systems, 2014, 39(2):435-462.

[24] 北京市环境保护局 . 北京市环境保护局 [EB/OL]. http://www.bjepb.gov.cn/.

[25] 河北省环境保护厅 . 河北省环境保护厅 [EB/OL]. http://www.hb12369.net/hjzw/hjjcyyj/jdxjc/.

[26] 环境保护部 . 大气污染物综合排放标准 [EB/OL]. [1997-01-01]. http://kjs.mep.gov.cn/hjbhbz/bzwb/dqhjbh/dqgdwrywrwpfbz/199701/t19970101_67504.htm.