Advances in Applied Mathematics 应用数学进展, 2017, 6(7), 801-807
Published Online October 2017 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2017.67097

Hans汉斯

Urban Air Quality Evaluation of Hubei Province Based on SOM Network and Association Rules

He Zhang, Bo Wang

School of Mathematical and Physics, China University of Geosciences, Wuhan Hubei
Email: 2390614186@qq.com

Received: Oct. 1st, 2017; accepted: Oct. 15th, 2017; published: Oct. 19th, 2017

Abstract

With the development of economy and the increase of population, all kinds of environmental problems have taken place in our city. Firstly, taking $PM_{2.5}$, PM_{10}, SO_2, CO, NO_2, and O_3 the six major air pollutants as indicators, self-organizing neural (SOM) network is used to study the urban air quality in Hubei province. The SOM neural network clustering model is established, and the results show that the air quality in Hubei province is reduced from the periphery to the center. Secondly, this paper uses the association rules in data mining and the classical Apriori algorithm to mine the correlation between $PM_{2.5}$, NO_2, and O_3 the three major air pollutants, the strong association rules are found finally.

Keywords

SOM, Neural Network, Association Rules, Air Quality, Hubei Province

基于SOM网络及关联规则的湖北省城市空气质量研究

张　贺，王　博

中国地质大学(武汉)数学与物理学院，湖北　武汉
Email: 2390614186@qq.com

收稿日期：2017年10月1日；录用日期：2017年10月15日；发布日期：2017年10月19日

摘　要

随着经济发展及人口增多，我们居住的城市陆续发生了各种环境问题。利用自组织竞争(SOM)神经网络

文章引用: 张贺, 王博. 基于SOM网络及关联规则的湖北省城市空气质量研究[J]. 应用数学进展, 2017, 6(7): 801-807.
DOI: 10.12677/aam.2017.67097

研究湖北省城市环境空气质量，以$PM_{2.5}$、PM_{10}、SO_2、CO、NO_2、O_3六个主要空气污染物作为指标，建立SOM网络模型，结果显示在空间分布上湖北省城市空气质量呈现从外围到中心降低的特征。采用数据挖掘中的关联规则，利用经典的Apriori算法挖掘主要空气污染物之间的关联关系，得到$PM_{2.5}$、NO_2、O_3之间的强关联规则。

关键词

SOM，神经网络，关联规则，空气质量，湖北省

1. 引言

空气与人们的生活密切相关，近年来，由于我国经济飞速发展，城市化进程不断加快，城市人口持续增长，人类活动使原本保持稳定的空气遭到越来越严重的污染。城市中大量的车辆和人口引发的汽车尾气以及居民生活和取暖等问题使空气质量开始逐渐恶化，严重威胁着城市居民的日常生活和人体健康，破坏城市生态。空气中含有的一些污染物如果超过一定浓度极有可能对人体产生健康威胁，引起各种各样的疾病。全面掌握城市空气污染源的排放数据，了解污染分布情况，完成对于空气质量监测、分析以及城市空气质量评价的研究，掌握城市空气质量在空间的分布，对城市规划与建设、污染控制、环境管理有重要的理论意义与实用价值。现今国内外都在进行环境质量评估相关工作，到现在为止，有很多评价空气质量的方法，比如模糊聚类法、层次分析法和人工神经网络法。空气本身涉及到众多的参数和复杂的数学模式，这些评价方法具有一定的局限性，很难对空气质量进行准确的评价。本文针对空气质量指数监测数据中存在的不确定性等问题，以 $PM_{2.5}$、PM_{10}、SO_2、CO、NO_2、O_3 六类主要空气污染物作为指标，建立自组织竞争网络模型(self-organizing feature map)，在保证数据的模糊性和随机性的基础上，对湖北省 13 个主要州市空气质量指数监测数据进行分析，得到湖北省城市空气质量在空间的分布情况。

在进行环境污染治理时，针对性地消除空气中的污染物、了解污染物之间的相关性有助于提高环境治理工作的效率。关联规则是描述数据库中的数据项(属性、变量)之间所存在的(潜在)关系的规则。2012 年我国环保部批准发布的《环境空气质量标准》(GB3095-2012)明确规定了 $PM_{2.5}$、PM_{10}、SO_2、CO、NO_2、O_3 六类主要污染物的浓度限制和分级标准[1]。本文以湖北省空气监测数据为依托，应用经典关联规则数据挖掘算法 Apriori 算法发现主要污染物之间的关联规则，找到他们的相关性。

2. 数据来源及数据预处理

选取 2015 年 1 月~2017 年 4 月湖北省 13 个州市(武汉市、黄石市、十堰市、宜昌市、襄阳市、鄂州市、荆门市、孝感市、荆州市、黄冈市、咸宁市、随州市、恩施自治州)6 个空气质量指标：$PM_{2.5}$、PM_{10}、SO_2、CO、NO_2、O_3，相应指标数据来自中国空气质量在线监测分析平台。

选取的六个指标具有不同的量纲和量纲单位，为了消除指标之间的量纲影响，采用最大最小法对数据进行标准化处理，使结果值映射到[0,1]之间。各指标平均值标准化数据见表 1。

3. SOM 网络模型在湖北省空气质量分析中的应用

芬兰学者 Kohonen 提出自组织竞争网络模型(SOM)，他认为，一个神经网络在接受外界输入模式时，

Table 1. Standardized data for average value of each index
表 1. 各指标平均值标准化数据

城市	$PM_{2.5}$均值	PM_{10}均值	SO_2均值	CO 均值	NO_2均值	O_3均值
武汉市	0.57	0.82	0.33	0.35	1.00	0.59
黄石市	0.54	0.70	0.65	0.81	0.43	0.32
十堰市	0.10	0.31	0.74	0.48	0.28	0.10
宜昌市	0.80	0.94	0.39	0.32	0.53	0.00
襄阳市	1.00	0.86	0.45	0.38	0.47	0.61
鄂州市	0.62	0.76	0.74	0.68	0.46	0.80
荆门市	0.57	1.00	0.86	0.18	0.51	0.46
孝感市	0.37	0.64	0.09	1.00	0.17	0.69
荆州市	0.71	0.97	1.00	0.45	0.54	0.83
黄冈市	0.21	0.31	0.16	0.64	0.23	1.00
咸宁市	0.09	0.35	0.00	0.34	0.00	0.83
随州市	0.49	0.68	0.19	0.70	0.17	0.81
恩施自治州	0.00	0.00	0.00	0.00	0.02	0.55

可以对输入信号特征进行自适应学习，从而自组织形成对输入模式将具有不同的响应特征的不同区域。在输出空间中，这些神经元形成一张功能相同的神经元靠得较近、功能不同的神经元分得较开的映射图，因此叫做自组织特征映射网络，通过竞争学习完成[2]。

当接受一种输入模式后，其输出层其中一个神经元得到最大限度的刺激从而竞争获胜，同时也因侧向相互作用使获胜神经元附近的一些神经元得到较大刺激。然后，修改这些神经元和输入神经元之间的连接权值，二维平面上获胜的输出神经元会随着输入模式改变发生相应改变。

SOM 网包括可以模拟感知外界输入信息的视网膜输入层以及模拟做出响应的大脑皮层两层，输出阵列见图 1。

本文根据自组织竞争网络具备的自组织、自适应的数据压缩、特征抽取等特点，以 $PM_{2.5}$、PM_{10}、SO_2、CO、NO_2、O_3 作为指标体系，建立 SOM 网络模型。应用 SOM 网络进行空气质量评价分类的步骤如下[3]：

1) 选取标准空气质量样本；

2) 对每一种标准空气质量样本进行学习，学习结束后，对具有最大输出的神经元标以记号；

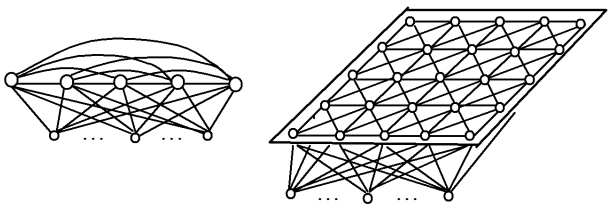
3) 将待分类数据输入到 SOM 模型中。

运用 MATLAB 软件编写程序，建立 SOM 网络，按照 SOM 算法步骤，设定学习率在前 1000 步训练中从 0.5 线性下降至 0.04，接着在训练到 10,000 步时减小至 0。优胜领域半径初值设定为 2 个节点，1000 个训练步时减至 0。然后将表中的数据样本进行归一化处理，输入网络并进行训练，依次进行 5、20、50、100 步数训练。当训练步数为 5 时，城市空气质量被分为 4 类，此时网络已经对数据进行了初步的分类。当训练步数为 50 时，分类更加细化，大多数都是单独被划分为一类，这时如果继续提高训练步数，已经不具备实际意义。根据实际情况选取 5 步训练结果，湖北省城市分类结果见表 2。

根据上述结果，以绿色颜色深浅表示城市的空气质量情况(绿色颜色深表示级别越高，城市空气污染度成都低，空气质量好)，绘制湖北省城市空气质量级别评价图，表征湖北省城市空气质量的空间分布特征，见图 2。

Table 2. Results of urban classification in Hubei province
表 2. 湖北省城市分类结果

级别	城市	$PM_{2.5}$ 均值	PM_{10} 均值	SO_2 均值	CO 均值	NO_2 均值	O_3 均值
I 级	恩施、咸宁	52.03	79.10	10.87	0.92	20.59	92.94
II 级	十堰	53.32	83.80	21.23	1.20	28.51	74.66
III 级	随州、孝感、黄冈、黄石	59.83	92.86	14.65	1.47	27.61	93.30
IV 级	武汉、襄阳、荆门、宜昌、荆州、鄂州	66.56	103.40	19.66	1.12	37.13	88.55



(a)一维线阵　　(b)二维平面线阵

Figure 1. The output array of SOM
图 1. SOM 网的输出阵列

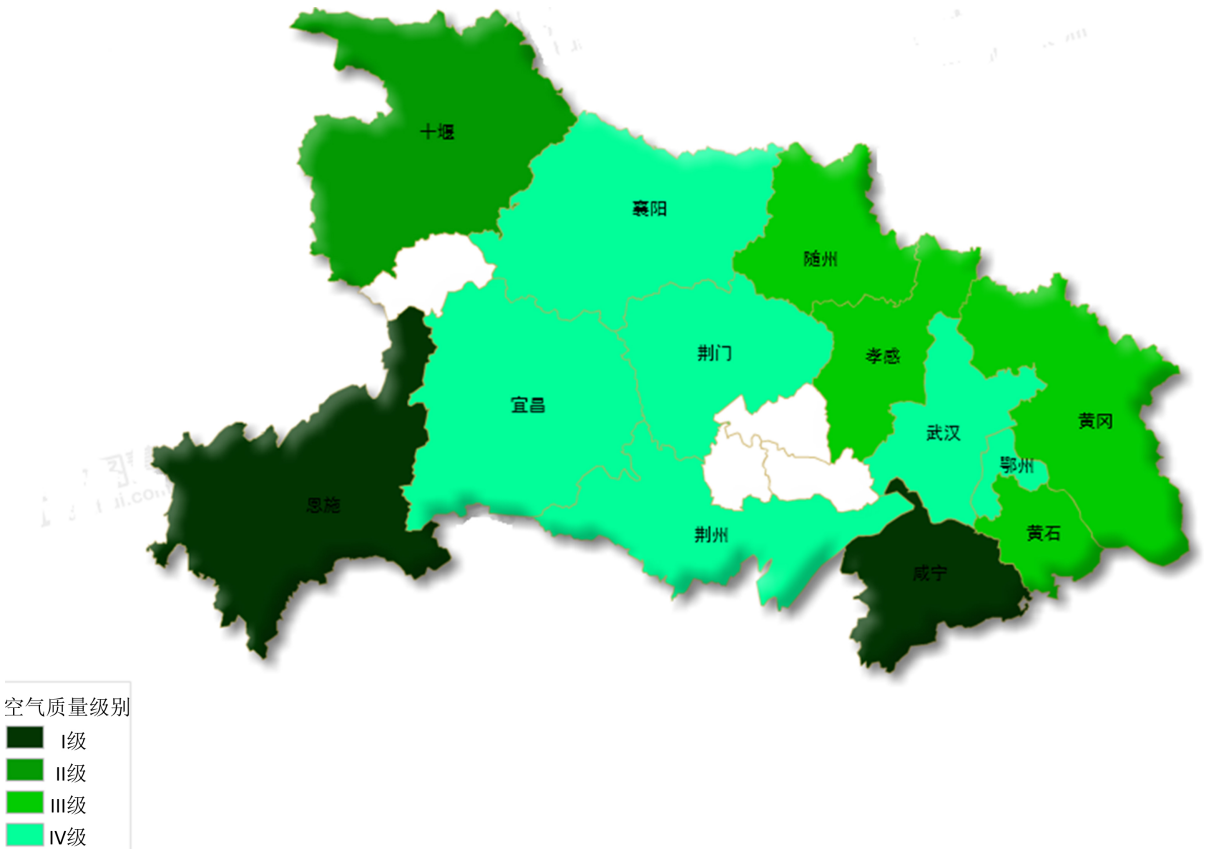


Figure 2. Urban air quality evaluation of Hubei province
图 2. 湖北省城市空间质量级别评价

由图可知，湖北省城市空气质量在空间上由外围向中心变差，空气污染程度加剧。绿色最深的城市是恩施自治州和咸宁市，分别位于湖北省西南部及东南部，空气质量最好；其次为十堰市，位于湖北省

西北部，空气质量较好；位于湖北省东北部的黄冈市、孝感市、黄冈市、黄石市绿色颜色较浅，城市空气质量较低，空气污染程度较低；荆门市、鄂州市、宜昌市、武汉市、荆州市、襄阳市位于湖北省内部地区，绿色最浅，是省内空气质量最差的区域。从基于 SOM 模型分析湖北省 13 个主要州市空气质量数据得到的结果来看，从总体来看，大部分城市都处在程度不同的环境污染问题当中，需要得到及时治理。

4. 关联规则挖掘空气污染物相关性

关联规则由 Agrawal 等人于 1993 年首先提出用于挖掘顾客交易数据库中项集间的联系，最成功的应用之一是超市的购物篮研究[4]。设数据集为 D，$I=\{i_1,i_2,\cdots i_m\}$ 为一个项目的集合，事务 T 表示项目子集 $(T\subseteq I)$，而每个子集都有唯一的标识 TID。X 是由项目子集构成的集合，叫做项集。当且仅当 $X\subseteq T$ 时，我们才认为事务 T 是包含项集 X 的。另外，当项集 X 包含 k 个项目时，项集 X 称为 k 项集[5]。

关联规则是形如 $X\Rightarrow Y$ 的逻辑蕴含式，其 $X\subset I, Y\subset I$ 中且 $X\cap Y=\varnothing$。而关联规则 $X\Rightarrow Y$ 指的就是事务数据库 D 的支持度，表示的是项集 X 在整个数据集 D 中所占的百分比[6]。支持度作为衡量关联规则强弱的重要标准，描述的是挖掘出来的规则在整个事务库中出现的频率。在挖掘过程中，用户可以根据实际需要来设定自身所需的支持度的阈值，通常称为最小支持度，记为 min_sup。

$$Support(X)=\left\|\{t\in D|X\subset t\}\right\|/\|D\| \tag{1}$$

$$Support(X\Rightarrow Y)=P(X\cup Y) \tag{2}$$

置信度是关联规则中对挖掘出来的关联规则正确率的评判标准，取值范围是[0,1]。与支持度一样，用户可以根据实际需要来设定相应的置信度阈值，称为最小置信度，记为 min_$conf$，其中 $X\subset I, Y\subset I, X\cap Y=\varnothing$。

$$Confidence(X\Rightarrow Y)=\frac{Sup(X\cup Y)}{Sup(X)}=\frac{P(XY)}{P(X)}=P(Y|X) \tag{3}$$

关联规则的挖掘算法中最经典的是 Apriori 算法[7]，Apriori 算法是 1994 年 Agrawal 提出的挖掘完全频繁项集中最具有影响力的算法。算法有两个关键的步骤：一是发现所有的频繁项集；二是生成强关联规则。该算法简单明了，易于实现，目前仍是使用最广泛的关联规则挖掘算法。关联规则按照变量类型分为布尔型关联规则和数值型关联规则，布尔型关联规则主要处理离散、种类化的数值类型，算法相对比较成熟。数值型属性的取值范围较广，在进行关联规则挖掘时通常将之转换成布尔型关联规则挖掘问题，即将属性取值划分为若干个区间，然后将每个区间映射为一个布尔型属性。

湖北省六种基本空气污染物监测数据为数值型数据，本文根据《环境空气质量标准》中规定的空气污染物浓度限值和分级标准(见表 3)，将指标数据映射为布尔型数据，得到 364 条记录，映射数据部分记录见表 4，各指标映射结果见图 3。

由上图可知，在记录中 PM_{10}、SO_2、CO 区间几乎没有变动，反映出这三种污染物在本文研究区域内基本相同，因此本文选取 $PM_{2.5}$、NO_2、O_3 指标，利用 Apriori 算法探寻这三种主要污染物的相关性。

Apriori 算法主要是通过产生一组更小的候选项集，根据阈值对产生的候选项集进行必要剪枝，以此来减少候选项集个数，最后由剩下的候选项集产生频繁项集，得到对用户有用的结果。运用 Python 软件编写程序，实现 Apriori 算法，设定 min_sup 为 50，min_$conf$ 为 0.8，最终得到以下关联规则：

规则 1：6-I&1-II→5-I　$conf$：0.8384；

规则 2：5-I&1-III→6-I　$conf$：0.8824；

规则 3：6-II&1-II→5-I　$conf$：0.9107；

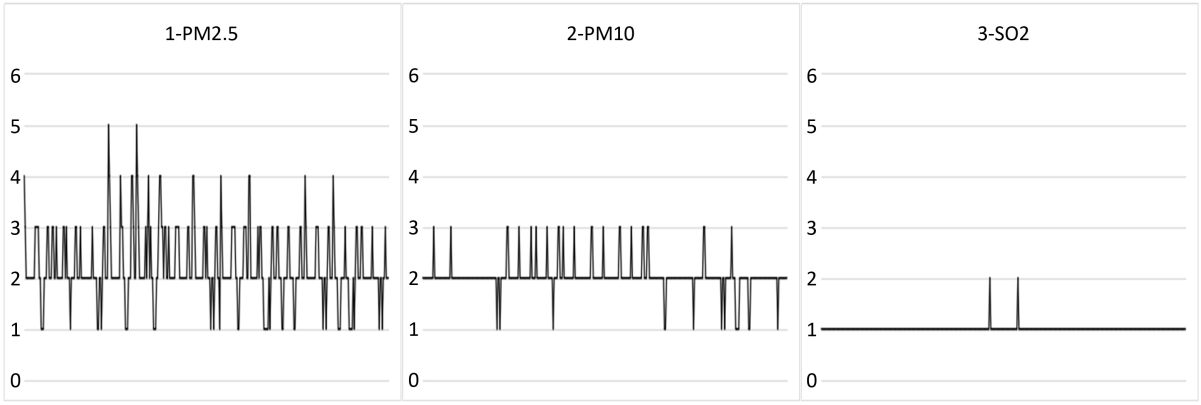
规则 4：5-II&1-III→6-I　$conf$：0.9545。

Table 3. Air pollutant concentration limits and classification standards
表 3. 空气污染物浓度限值和分级标准(单位：mg/m^3)

$PM_{2.5}$.	PM_{10}	SO_2	CO	NO_2	O_3	污染物级别划分
≤0.035	≤0.050	≤0.050	≤2.000	≤0.040	≤0.100	I 级
0.035~0.075	0.050~0.150	0.050~0.150	2.000~4.000	0.040~0.080	0.100~0.160	II 级
0.075~0.115	0.150~0.250	0.150~0.475	4.000~14.00	0.080~0.180	0.160~0.215	III 级
0.115~0.150	0.250~0.350	0.475~0.800	14.00~24.00	0.180~0.280	0.215~265	IV 级
0.150~0.250	0.350~0.450	0.800~1.60	24.00~26.00	0.280~0.565	0.265~0.800	V 级
≥0.250	≥0.420	≥1.60	≥36.00	≥0.565	≥0.800	VI 级

Table 4. Partial record of mapping data
表 4. 映射数据部分记录

记录	$PM_{2.5}$	PM_{10}	SO_2	CO	NO_2	O_3
1	1-IV	2-II	3-I	4-I	5-II	6-I
2	1-III	2-II	3-I	4-I	5-II	6-I
3	1-II	2-II	3-I	4-I	5-II	6-I
4	1-II	2-II	3-I	4-I	5-II	6-II
5	1-II	2-II	3-I	4-I	5-II	6-II
6	1-II	2-II	3-I	4-I	5-I	6-II
7	1-II	2-II	3-I	4-I	5-I	6-II
8	1-II	2-II	3-I	4-I	5-I	6-II
9	1-II	2-II	3-I	4-I	5-II	6-II
10	1-II	2-II	3-I	4-I	5-II	6-II



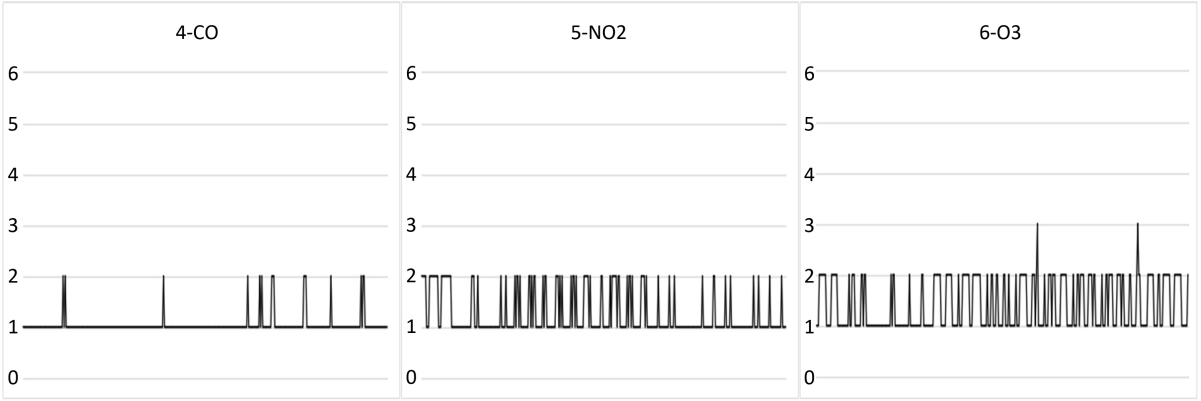


Figure 3. Mapping results of each index
图 3. 各指标映射结果

分析规则我们发现 $PM_{2.5}$、NO_2、O_3 这三种污染物之间存在很强的相关性，当 O_3I 级、$PM_{2.5}$I 级时，NO_2 为 I 级的置信度为 83.84%；当 NO_2I 级、$PM_{2.5}$III 级时，得到 O_3 为 I 级的置信度为 88.24%；当 O_3 为 II 级、$PM_{2.5}$II 级时，NO_2 为 I 级的置信度为 91.07%；当 NO_2II 级、$PM_{2.5}$III 级时，O_3 为 I 级的置信度为 95.45%。

5. 结论

1) 利用神经网络中的自组织(SOM)神经网络进行湖北省 13 个州市空气质量评价，建立 SOM 网络模型，将 13 个州市分为四类，发现在空间分布上湖北省呈现从外围到中心空气质量降低的特征。评价结果与实际相符，从实践上证明了神经网络用于城市空气质量评价的优越性，SOM 模型的建立解决了 BP 神经网络的精度低的缺点，大大提高了数据分类的精度，为人工神经网络在城市空气质量评价领域提供了新的途径。

2) 应用经典关联规则 Apriori 算法发现 $PM_{2.5}$、NO_2、O_3 三种主要污染物之间的强关联规则，找到他们的相关性，有助于我们了解污染源，以便制定相应环境治理措施。

参考文献 (References)

[1] 环境保护部环境监测司. “十二五”环境监测工作手册[M]. 北京: 中国环境科学出版社, 2012: 300-301.

[2] 方晗. 稻谷干燥模型数学模拟与神经网络仿真[D]: [硕士学位论文]. 沈阳: 东北大学, 2004: 16-17.

[3] 王小川, 史峰. MATLAB 神经网络 30 个案例分析[M]. 北京: 北京航空航天大学出版社, 2010: 159-169.

[4] Agrawal, R., Imielinski, T. and Swami, A. (1993) Mining Association Rules between Sets of Items in Large Databases. *Proceedings of the* 1993 *ACM SIGMOD International Conference on Management of Data*, Washington, D.C., 25-28 May 1993, 207-216. https://doi.org/10.1145/170035.170072

[5] 李雄飞, 董元方, 李军, 等. 数据挖掘与知识发现[M]. 北京: 高等教育出版社, 2010.

[6] 刘小珊, 罗文强, 李飞翱, 黄丽. 基于关联规则的滑坡演化阶段判识指标[J]. 地质科技情报, 2014, 33(2): 119-123.

[7] Chang, K.C.-C., He, B., Li, C.K., *et al*. (2004) Structured Database on the Web: Observations and Implications. *SIGMOD Record*, **33**, 61-70. https://doi.org/10.1145/1031570.1031584

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2324-7991，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：aam@hanspub.org