

基于分层递阶粒度聚类法的空气质量评价^{*}

汪小寒¹, 张燕平², 赵 姝², 张 铃²

(1. 安徽师范大学 数学计算机科学学院, 安徽 芜湖 241003; 2. 安徽大学 计算机科学与技术学院, 合肥 230039)

摘 要: 针对气候变化与空气质量评价问题, 考虑到空气质量数据具有复杂、不确定、不完备、海量高维的特点, 提出基于分层递阶商空间的粒度聚类分析法。利用分层递阶模型, 通过属性函数对原问题的论域进行划分, 获得问题求解的不同层次粒度空间, 可以将不确定信息转换成确定性信息, 降低问题复杂性, 易推广至高维海量数据分析中。宣城市空气质量评价实验结果表明, 这种方法能很好地反映该市空气质量的日变化和季节变化, 进一步验证了其有效性。

关键词: 粒度; 商空间; 分层递阶; 聚类; 空气质量

中图分类号: TP391

文献标志码: A

文章编号: 1001-3695(2013)01-0192-03

doi:10.3969/j.issn.1001-3695.2013.01.049

Air quality evaluation based on delaminated granular clustering method

WANG Xiao-han¹, ZHANG Yan-ping², ZHAO Shu², ZHANG Ling²

(1. School of Mathematics & Computer Science, Anhui Normal University, Wuhu Anhui 241003, China; 2. School of Computer Science & Technology, Anhui University, Hefei 230039, China)

Abstract: Due to the air quality evaluation of climate change has complex, uncertain, imperfect, high-dimension and massive data properties, this paper put forward the granular clustering method based on delamination. New quotient space could be obtained after universe partition by attribute function based on delamination in the theory of quotient space. And the new space could make complex and uncertain problem simple and certain. Experimental result of air quality evaluation on Xuancheng city also shows that this method is more effective to its day and season's change.

Key words: granula; quotient space; delamination; cluster; air quality

随着我国工业化和城镇化快速发展, 大气污染问题日益严重^[1], 对生态环境和人类健康危害较大。为了控制和减少污染, 就必须采用科学有效的方法对大气污染进行监测和统计分析。因此, 大气质量分析与评价是环境科学和人工智能领域的热门交叉课题^[2]。环境空气质量除了直接受局地大气污染物排放影响, 还受到区域和全球气候变化的影响。针对气候变化引起的环境空气质量状况的分析可以准确地为预报空气质量、治理大气污染、降低环境对健康的损害提供科学依据。而我国在涉及气候变化对环境空气质量影响方面的研究还处于起步阶段, 对气候变化影响环境空气质量的研究相对较少^[3]。现有的大气质量典型评价方法是空气污染指数法(即API法)。API法将常规监测的几种空气污染物的浓度简化成为一个数值形式, 并分级表征空气质量状况, 该方法简单直观, 但空气质量状况很难仅用一个简单的数据说明清楚, 其硬性规定成分过重, 不能全面反映实际情况。其他的主要方法有, 高斯模型^[1]、灰色聚类方法^[4]、主成分分析法^[5]等。这些方法取得了很好的应用效果, 但这些模型和方法未考虑气候因素对空气质量的影响, 大多是基于少量数据分析。例如, 文献[1]选用了10组检测数据进行分析; 文献[4]对2005—2009年共五年的三种颗粒物含量的年平均值数据进行聚类分析; 文献[5]对

2004—2008年中的七个属性年平均值进行分析。由于空气质量数据具有复杂、不确定、海量高维和非线性等特点, 现有的模型和方法不能很好地满足这类数据分析的需要。要实现对大气质量数据的精确分析必然依赖有效的问题求解^[6], 分层递阶粒度模型为此提供了有力的分析工具。

1 商空间基本理论

商空间理论^[7~12]用一个三元组 (X, f, T) 描述一个问题, 其中 X 是论域, f 是论域的属性, T 是论域的结构。划分是构成不同粒度世界的方法, 问题的不同粒度是对论域进行不同的划分, 对应于不同的等价关系 R 。利用等价关系 R 对论域进行划分, 可以将原问题细粒度空间转换成新的粗粒度商空间。粒度计算是对问题在各种商空间上的分析。论域划分方法主要有投影划分法、属性划分法、结构划分法^[8]。本文空气质量数据分析就是采用属性函数划分法获得新的粒度和问题空间。

1.1 属性函数划分法

属性划分法将属性相同或相似的元素归为一类。设 $f: X \rightarrow Y$ 是属性函数, 若 f 是单值的, 则可利用 f 来定义划分。一般来说, 对 Y 的结构比较清楚, 如 Y 是实数集或 R^n 欧式空间,

收稿日期: 2012-06-12; **修回日期:** 2012-07-26 **基金项目:** 国家自然科学基金资助项目(61175046, 61073117); 安徽高校省级自然科学基金项目(KJ2012Z121); 安徽师范大学人才培育基金资助项目(2010rcpy037)

作者简介: 汪小寒(1978-), 女, 安徽枞阳人, 讲师, 硕士, 主要研究方向为智能计算及其应用(hanxiaohu@sina.com); 张燕平(1962-), 女, 安徽巢湖人, 教授, 博士, 主要研究方向为智能计算; 赵姝(1979-), 女, 安徽巢湖人, 副教授, 博士, 主要研究方向为智能计算; 张铃(1937-), 男, 福建福清人, 教授, 主要研究方向为人工智能。

可以利用 Y 的分类来定义 X 中对应的分类。方法为: 设 $\{Y_i\}$ 是 Y 的一个划分, 定义 $X_i = \{x | f(x) \in Y_i\}$, 则 $\{X_i\}$ 是 X 的一个划分。

1.2 分层递阶方法

文献[8]中指出人类智能可以从极不相同的粒度或者尺度上观察和分析同一问题, 而且还能够很快地从一个粒度世界跳到另一个粒度世界, 往返自如。分层递阶方法就是粒度由粗到细, 从全局到局部的解决问题的方法。

文献[13]指出分层递阶结构是模糊问题和清晰问题之间的一座桥梁。人类对客观世界中的模糊与清晰、不确定与确定的认识是相对的, 在一定程度上可以相互转化。

2 空气质量分层递阶粒度聚类分析法

2.1 空气质量评价分层递阶粒度聚类分析法

空气质量评价的分层递阶粒度聚类分析法主要有三步:

a) 用商空间理论中的划分方法对论域进行划分, 获得空气质量数据某一个层次上的粒度空间。

b) 在上述粒度层次空间中用聚类方法分类。

c) 对分类结果进行分析、综合, 获得不同层次上的确定性信息, 重复 a)b), 直到获得令人满意的聚类结果。

2.2 空气质量评价中的分层递阶粒度模型分析

用商空间模型分析空气质量数据, 论域 X 由气象信息数据和 PM10 值按照日期顺序构成集合。论域属性 f 的主要构成如下: $f_1(x)$ 为日平均气压, $f_2(x)$ 为日平均气温, $f_3(x)$ 为日平均相对湿度, $f_4(x)$ 为日平均总云量, $f_5(x)$ 为日平均低云量, $f_6(x)$ 为日降水量, $f_7(x)$ 为前一天降水量, $f_8(x)$ 为日平均风速, $f_9(x)$ 为前一天 PM10 值, $f_{10}(x)$ 为当日 PM10 值。这样 $f = (f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8, f_9, f_{10})$ 。结构 T 表示各个属性元素之间的内在关系, 当论域 X 和属性函数 f 确定之后, 结构 T 也就随之确定。

用分层递阶粒度模型来分析空气质量数据, 是用时间属性对论域 X 进行划分。分别用 2 日、5 日、15 日三种不同的时间属性函数对论域进行划分, 可得出三种不同层次的粒度空间, 记为 $([X]_1, [f]_1, [T]_1), ([X]_2, [f]_2, [T]_2), ([X]_3, [f]_3, [T]_3)$, 实现了粒度的粗细变换。例如, 在 $([X]_1, [f]_1, [T]_1)$ 细粒度空间中可以获得空气质量的细微变化信息, 如日变化情况; 而在 $([X]_3, [f]_3, [T]_3)$ 粗粒度空间中可以获得空气质量改变的总体变化信息, 如季节性变化情况。

利用分层递阶粒度模型获取不同层次商空间后, 在某个层次上模糊的信息在另一个层次上会转变为清晰的信息, 在某个层次上不确定信息在另一个层次上可以转变为确定性的信息, 因此这种分层递阶的方法更适合具有不完备、不确定性数据的分析。另外, 通过分层递阶方法可以逐步地达到问题求解的目的, 降低问题复杂性, 这种方法更容易推广到高维海量数据的分析之中。

3 基于分层递阶粒度聚类方法宣城市空气质量分析实验

3.1 实验数据介绍

实验数据源于宣城市气象局和宣城市环境监测保护中心,

仿真实验通过 SPSS 工具软件模拟执行。宣城地区气候具有春暖、夏热、秋爽、冬寒的特点, 四季分明, 一年中夏季约 121 天、冬季约 102 天、春季约 73 天、秋季约 69 天。夏季高温多雨, 冬季寒冷少雨, 降水集中在暖热季节。根据宣城市的季节情况, 实验中的冬季为 1、2、12 月, 春季为 3、4 月, 夏季为 5~9 月, 秋季为 10、11 月。

实验数据选取了宣城市 2003 年的气候因子和 PM10 值进行分析, 选取的具体数据项为日平均气压、日平均气温、日平均相对湿度、日平均总云量、日平均低云量、日降水量、前一天降水量、日平均风速、前一天 PM10、当日 PM10, 共十个属性值组成一个样本。原始样本数据按照连续的日期顺序排列, 实验中通过属性函数划分法对原始样本数据取均值, 获得三种不同大小的粒度, 分别记为粒度 1、粒度 2、粒度 3。实验中的聚类方法选用的是 K-means 方法, 实验中均设置聚类值为四类。

实验 1 选取粒度 1, 用时间属性 2 天对原始数据进行划分, 全年样本共 179 个。

实验 2 选取粒度 2, 用时间属性 5 天对原始数据进行划分, 全年样本共 79 个。

实验 3 选取粒度 3, 用时间属性 15 天对原始数据进行划分, 全年样本共 25 个。

为了便于表达, 实验中将样本按照月份和日期进行编号, 如 101。实验 1 中是由原始样本 4 月 7 日和 4 月 8 日数据得到的; 实验 2 中是由原始样本 1 月 1 日~1 月 5 日共五天的数据得到的; 实验 3 中的 101 是由原始样本 1 月 1 日~1 月 15 日共 15 天的数据得到的。

3.2 实验结果

实验 1~3 的聚类结果如表 1~3 所示。三个实验的 F 检验值和 Sig 值的对照如表 4 所示。聚类结果中每类样本数量情况统计如图 1、2 所示。

表 1 实验 1 聚类结果

类别	样本分布情况	个数
1	329, 407, 413, 415, 417, 501, 515, 517, 519, 521, 523, 525, 527, 529, 601, 603, 609, 615, 617, 619, 621, 623, 625, 701, 703, 709, 715, 717, 721, 723, 725, 727, 729, 801, 803, 805, 807, 809, 811, 813, 815, 817, 819, 821, 823, 825, 827, 829, 901, 903, 905, 907, 913, 915, 917, 919, 921, 923, 925, 927, 929, 1001, 1009, 1101, 1105	65
	119, 125, 201, 207, 209, 211, 213, 215, 217, 221, 223, 225, 227, 301, 303, 305, 313, 315, 317, 401, 403, 409, 411, 419, 423, 425, 427, 429, 503, 511, 513, 909, 1003, 1005, 1007, 1011, 1013, 1107, 1109, 1117, 1119, 1125, 1127, 1201, 1205, 1209	
2	101, 103, 105, 107, 109, 111, 113, 115, 117, 121, 123, 127, 129, 203, 205, 219, 307, 309, 311, 319, 321, 323, 325, 327, 405, 421, 509, 1015, 1017, 1019, 1021, 1023, 1025, 1027, 1029, 1103, 1111, 1113, 1115, 1121, 1123, 1129, 1203, 1207, 1211, 1213, 1215, 1217, 1219, 1221, 1223, 1225, 1227, 1229	46
	505, 507, 605, 607, 611, 613, 627, 629, 705, 707, 711, 713, 719, 911	
3		54
4		14

表2 实验2 聚类结果

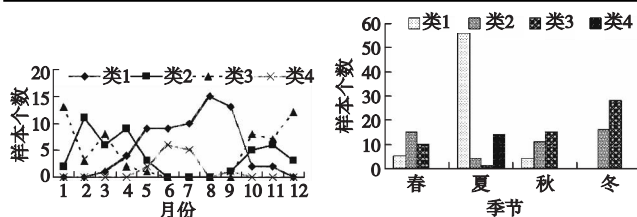
类别	样本分布情况	个数
1	101, 106, 111, 116, 121, 126, 201, 306, 316, 1111, 1121, 1211, 1216, 1221, 1226	15
2	321, 326, 406, 411, 601, 921, 926, 1016, 1021, 1026, 1101	11
3	206, 211, 216, 221, 301, 311, 401, 416, 421, 426, 1001, 1006, 1011, 1106, 1116, 1126, 1201, 1206	18
4	501, 506, 511, 516, 521, 526, 606, 611, 616, 621, 626, 701, 706, 711, 716, 721, 726, 801, 806, 811, 816, 821, 826, 901, 906, 911, 916	27

表3 实验3 聚类结果

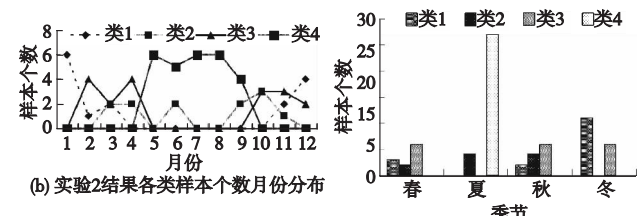
类别	样本号	个数
1	101, 116, 201, 1201, 1216	5
2	316, 916, 1016	3
3	216, 301, 401, 416, 1001, 1101, 1116	7
4	501, 516, 601, 616, 701, 716, 801, 816, 901	9

表4 实验结果中 F 值和 Sig 值的对照

名称	实验1		实验2		实验3	
	F 值	Sig 值	F 值	Sig 值	F 值	Sig 值
$f_1(x)$	138.256	0.000	91.223	0.000	46.388	0.000
$f_2(x)$	159.659	0.000	104.365	0.000	41.657	0.000
$f_3(x)$	64.356	0.000	29.899	0.000	5.532	0.006
$f_4(x)$	39.081	0.000	23.377	0.000	7.342	0.002
$f_5(x)$	35.591	0.000	15.105	0.000	4.828	0.011
$f_6(x)$	39.726	0.000	5.657	0.002	2.033	0.142
$f_7(x)$	28.417	0.000	1.340	0.269	2.329	0.105
$f_8(x)$	3.409	0.019	3.287	0.026	2.500	0.089
$f_9(x)$	4.991	0.002	3.274	0.026	1.308	0.299
$f_{10}(x)$	5.348	0.002	2.737	0.050	1.258	0.316



(a) 实验1结果各类样本个数月份分布 (a) 实验1结果各类样本个数季节分布



(b) 实验2结果各类样本个数月份分布 (b) 实验2结果各类样本个数季节分布

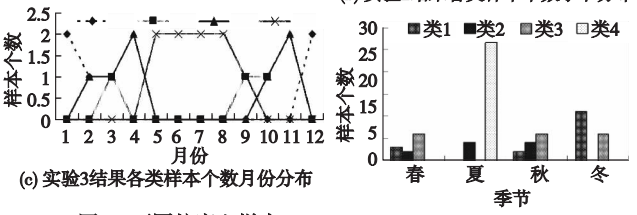


图1 不同粒度上样本个数月份分布对照

(c) 实验3结果各类样本个数月份分布 (c) 实验3结果各类样本个数季节分布

图2 不同粒度上各类样本个数季节分布对照

3.3 实验结果分析

图1结果显示了各类样本数量在各个月份的分布状况。图2结果显示了各类样本数量随季节的变化情况:(a)中春、秋季有三种样本,夏季有四种样本,冬季有两种样本,日变化较

为明显,季节变化不明显;(b)中夏季由四种样本变成两种样本,聚类效果更加明显;(c)中春、夏、秋、冬均只有两种样本,夏、冬季中类别区分度大,聚类效果更佳,最能反映宣城市空气质量的季节性变化特征。因此,这种分层递阶的方法可以使原来在细粒度1商空间中表现为不确定的信息在粗粒度3商空间中转换为确定的信息。

表4中数据表明日平均气压、日平均气温对分类影响最大;在粗粒度3的商空间中聚类结果受气象条件影响大,受PM10值的影响不大;在细粒度1空间中聚类结果不仅受各种气候因素影响,而且受PM10影响较大。这种从多个层次商空间中获取有效信息的方法,能够满足不同问题分析的需要。

实际上,宣城市近地面层空气污染与气象条件密切相关,存在着明显的日变化和季节变化,实验结果与事实情况基本一致,这也进一步验证了本文方法的有效性。实验中样本是没有经过任何处理的原始数据,因此结果更加准确、可靠。将原问题空间转换成不同层次粒度空间的方法,能更好地满足问题求解的需要,降低计算复杂性,易于推广到高维海量数据分析中。

4 结束语

本文提出空气质量评价的分层递阶粒度聚类分析方法,通过分层递阶方法,将原问题空间分解成不同层次粒度空间,在新的商空间中解决问题,可获得更多的有效信息,可以使某个粒层上的不确定性信息转换成另一个粒层上的确定性信息。这种分层递阶的方法可以降低问题复杂性,提高效率,更加符合人类解决复杂问题的思维方式。将这种方法应用到宣城市空气质量数据分析实验中,实验结果既能在粗粒度空间中反映宣城市实际污染季节性变化特征,又能在细粒度空间中体现宣城市空气质量的日变化情况。分层递阶粒度模型下的大气质量评价与其他更加高效算法的结合正在进一步研究之中。

参考文献:

- [1] 黄金杰,杨桂花,马骏驰. 基于高斯的大气污染评价模型[J]. 计算机仿真, 2011, 28(2): 101-104.
- [2] 王守权. 大气质量评价与预测的智能方法研究[D]. 长春: 吉林大学, 2011.
- [3] 孙家仁, 许振成, 刘煜. 气候变化对环境空气质量影响的研究进展[J]. 气候与环境研究, 2011, 16(6): 805-814.
- [4] 边敏娟, 钱骏, 刘志红. 基于灰色聚类法的成都市大气污染评价[J]. 四川环境, 2011, 30(5): 59-62.
- [5] 付强, 王涛, 刘德惠. 主成分分析在环境空气质量评价中的应用[C]//中国环境科学学术年论文集. 北京: 中国环境科学出版社, 2011: 1183-1185.
- [6] 王玉平, 张志跃. 一维大气污染模型的特征——块中心差分法[J]. 计算力学学报, 2011, 28(5): 693-698.
- [7] 李道国, 苗夺谦, 张红云. 粒度计算的理论与方法[J]. 复旦大学学报, 2004, 43(5): 837-841.
- [8] 张钹, 张铃. 问题求解的理论与应用[M]. 北京: 清华大学出版社, 2007.
- [9] 张燕平, 张铃, 吴涛. 不同粒度世界的描述法——商空间法[J]. 计算机学报, 2004, 27(3): 41-45.
- [10] 张铃, 张钹, 张燕平. 模糊度的结构分析[J]. 中国科学: 信息科学, 2011, 41(7): 820-825.
- [11] 张铃, 张钹. 模糊相容商空间与模糊子集[J]. 中国科学: 信息科学, 2011, 41(1): 1-11.
- [12] 李刚, 万幼川, 管玉娟. 应用商空间理论的遥感影像多粒度合成分割[J]. 应用科学学报, 2011, 29(4): 390-395.
- [13] 王国胤, 张清华, 马希骞. 知识不确定性问题的粒计算模型[J]. 软件学报, 2011, 22(4): 676-670.