海量大气颗粒物成分分析系统①

张梦瑶[1,2], 廉东本[2], 赵 奎[2], 马元婧[2]

[1](中国科学院大学, 北京 100049)
[2](中国科学院沈阳计算技术研究所, 沈阳 110168)

摘 要: 2011 年以来, 我国多地出现了雾霾天气, 对大气颗粒成分分析有助于人们了解雾霾形成的原因, 制订有效的应对措施. 本文的主要目的是对于大气颗粒物成分进行命名. 传统颗粒物的命名是在经验的基础上, 对颗粒进行逐个的命名. 若将该过程自动化, 难点有两个: 数据规模太大、人工经验难以量化. 本文使用数据挖掘的工具, 首先进行了一次聚类分析, 降低了数据规模. 为了解决人工经验难以量化的问题, 使用逻辑回归分类算法, 并进行了调优, 使正确率达到了业务处理的要求.

关键词: 单颗粒; 聚类; 神经网络; 分类; 逻辑回归

Massive Atmosphere Particulate Matter Analysis System

ZHANG Meng-Yao[1,2], LIAN Dong-Ben[2], ZHAO Kui[2], MA Yuan-Jing[2]

[1](University of Chinese Academy of Sciences, Beijing 100049, China)
[2](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)

Abstract: Since 2011, fog and haze has been appeared in many places of China. Atmosphere particulates components analysis could help people to know the reason of the fog and haze so that we can take measures to deal with it. The purpose of this system is to name the atmosphere particulates, which means to classify the atmosphere particulates into seven common types. In traditional practice, we need to name the particulates artificially and one by one. But, if we want to do this automatically, there are two difficulties we need to solve. First, the scale of data is large. Secondly, it is hard to make rules to summarize the human experience. In this system, we try to use data mining technology to solve the two problems. To decrease the scale of data, we use the Adaptive Resonance Neural Cluster Algorithm, and to summarize the human experience, we use the Logistic Regression Classification Algorithm. At last, we adjust the models to get a better accuracy and meet the needs of the actual application.

Key words: single particle aerosol; cluster; neuron network; classification; logistic regression

城市大气颗粒可分为粉尘、烟尘和雾滴, 其中粉尘与雾结合可能形成相对稳定的“霾”, 对区域气候产生较大影响[1]. 分析其成分的组成可以帮助我们了解环境状况及环境恶化的原因, 为环境保护部门出台行之有效的措施提供依据, 而该系统的目的也正是对颗粒物的成分进行分析.

该系统的主要目的是将所有的空气颗粒划分为常见的七种类型: 矿物质(M)、重金属(HM)、大分子有机物(HOC)、有机碳(OC)、元素碳(EC)、钠钾(NaK)、钾(K).

目前对于空气颗粒的分析, 在硬件上可以使用单颗粒分析技术, 收集到颗粒级别的详细信息[1], 但是后续对于颗粒的识别工作目前基本需要人工完成. 由于颗粒物规模很大(可达到每天 20 万颗粒), 人工完成所有颗粒命名不现实, 所以较常采用样本估计的方法, 准确率不足.

本系统采用数据挖掘的方法, 使用聚类降低数据的规模, 使用分类方法完成颗粒的自动命名, 在保证正确率的前提下, 大大减少了人工的工作量.

① 收稿时间:2015-01-15;收到修改稿时间:2015-03-18

1　技术概述

1.1　大气颗粒物研究方法

目前大气颗粒物成分的分析方式主要为总体分析和单颗粒分析.

总体分析是以采集样本的总体作为研究对象. 该方法使用光谱或酸碱等方法对样本进行分析, 可以分析出样本整体表现出的光谱或化学特征, 进而推理出样本所含成分. 但精确程度较低.

单颗粒分析法以空气颗粒为分析单位, 可以对颗粒的形貌、粒度等单体性质做分析, 精确程度较高. 本系统中质谱仪使用的分析方法即为单颗粒分析法. 该质谱仪可以测算每个颗粒包含离子的质荷比及质谱信息.

1.2　颗粒质谱数据聚类分析

质谱数据分析中常用的聚类算法为: 等级方法、分区方法、基于密度的方法. 等级分类方法常见一个以层次分解给定的颗粒数据集合; 分区分类方法创建几个以某点为中心的分区, 然后改善分区的质量; 基于密度的方法是指聚类是根据密度情况来划分的, 如自然划分等.

在本系统中使用的共振神经网络算法属于基于分区的算法. 该算法的特点是可以快速识别新输入的模式类型, 即由聚类中心所代表的模式类. 并且能够根据环境输入, 自动将不同于已知模式类, 但又具有相似特征的输入归为新的一类, 并将学习到的这个新类的聚类中心保存下来, 作为一种新的模式类参与以后的分类.

1.3　逻辑回归模型

逻辑回归算法是一个主要解决 0–1 问题的分类算法. 若属于指定分类, 则对于该分类问题的结果为 1, 不属于则结果为 0. 其概率假设为 sigmoid 函数, 如图 1.

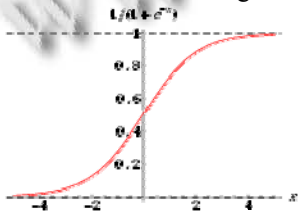


图 1　sigmoid 函数图像

其概率在 0–1 之间, 当样本足够大时, 可以有良好的分类表现.

相比较其他的分类算法, 逻辑回归算法的优势在于:

① 收敛速度快.

② 不需要满足严格的条件假设

③ 逻辑回归算法有很好的概率解释. 如与 某种颗粒相关性较大的系数有很大的可能也较大.

2　系统设计

2.1　系统目标

该系统的主要目标是完成对质谱仪收集的大气颗粒物成分命名, 即将颗粒物分类到七种常见的颗粒物类别中去, 这七种颗粒物分别为: 矿物质(M)、重金属(HM)、大分子有机物(HOC)、有机碳(OC)、元素碳(EC)、钠钾(NaK)、钾(K).

2.2　模块划分

如图 2, 根据颗粒物的处理流程, 该系统分为五个主要步骤: 质谱仪收集并电离大气颗粒、质谱仪数据的预处理、质谱聚类分析、自动命名、大气颗粒物成分统计.

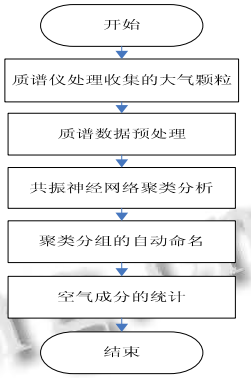


图 2　系统流程图

2.3　模块概述

(1)质谱仪收集并电离空气颗粒.

该步骤是系统的数据基础. 质谱仪使用单颗粒分析技术, 在定点收集某个地区、某个时间段内的空气颗粒, 并将空气颗粒电离为离子. 在电离过程中会产生离子图谱. 后期处理的数据即为颗粒所含离子及其图谱信息. 该过程质谱仪的输入为收集到的空气颗粒, 输出为空气颗粒对应的离子及离子质谱信息.

(2)质谱数据的预处理

该步骤主要是对后面步骤的数据准备. 其作用是将属于同一颗粒的离子聚集起来, 并选取合适的用于聚类的特征. 离子的质谱信息中包含峰高、峰面积、

相对峰面积等, 可根据需要选择一个特征, 并形成颗粒信息图谱.

(3)质谱聚类分析

该步骤主要是为了降低数据处理的规模. 使用共振神经网络算法聚类算法, 将相似度达到一定程度的颗粒聚集到一个分组中, 在后面的分类中, 就可以以组为单位运算.

(4)颗粒物自动命名

该步骤的输入为聚类分析分组结果的平均数据, 输入为每组的颗粒名称. 其核心算法使用逻辑回归的分类算法, 使用少部分颗粒进行人工训练, 其余颗粒进行验证.

(5)成分的统计

颗粒物成分被命名后, 为了对环境状况进行宏观的评估, 需要对空气中颗粒物的组成进行统计, 以便于制定应对措施.

3 系统模块实现

3.1 数据预处理

质谱仪的功能是收集空气中的大气颗粒, 对其进行电离, 每个颗粒会被电离为多个带有电荷的离子. 质谱仪中存储的离子信息包含离子所属颗粒的编号、颗粒被收集的时间、地点等. 带有电荷的离子会在质谱仪中形成一个频谱图像, 质谱仪可以记录该频谱图像形成的峰面积以及峰高. 根据每个离子在机器中的飞行时间, 可以测算出离子的质荷比(离子的质量与所带电荷之比). 即质谱仪通过分析可以得到每个颗粒中所含离子的质荷比及其形成频谱的峰高、峰面积.

数据预处理首先将具有同一个颗粒编号的离子聚集起来, 根据质荷比大小进行排序, 其在数据库中的存储格式如图 3 所示. 其中 OFMASSID 表示颗粒编号, MTC 表示质荷比, PEAKAREA 表示峰面积, RELAREA 表示相对峰面积, PEAKHEIGHT 表示峰高.

OFMASSID	MTC	PEAKAREA	RELAREA	PEAKHEIGHT
MASS-No_1-73558565210648	-106	69.000000	0.010700	6.000000
MASS-No_1-73558565210648	-99	178.000000	0.027600	17.000000
MASS-No_1-73558565210648	-98	41.000000	0.006400	6.000000
MASS-No_1-73558565210648	-96	3549.000000	0.550800	255.000000

图 3 离子信息

此时数据库中存储的关键离子属性有三个: 峰高、峰面积、相对峰面积, 而后期的处理则只需依据一个属性作为判别标准. 此时可根据需要选择一个属性, 如选择峰面积, 则每个离子所带有的信息为质荷比和峰面积. 每个颗粒可绘制出如图 4 的质谱图.

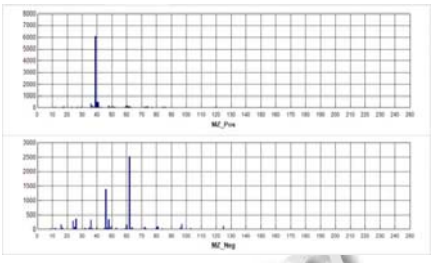


图 4 单个颗粒图谱

该质谱图的横坐标表示颗粒所含离子的质荷比, 纵坐标表示离子的峰面积. 其中第一幅图表示的是颗粒中所含的质荷比大于 0 的离子信息(即离子所带的电荷为正), 第二幅图表示的是颗粒中所含的质荷比小于 0 的离子信息(即离子所带的电荷为负).

3.2 质谱数据的聚类分析算法

对于气溶胶颗粒的聚类使用的是共振神经网络算法. 其算法的过程如下:

共振神经网络算法

repeat:

1: 对所有的颗粒数据进行规约, 将顺序打乱后依次输入.

2: 将第一个输入的颗粒作为初始中心.

3: 新输入的颗粒与所有中心计算余弦相似度, 并根据计算结果选择加入哪个分组或自己创建分组.

until 颗粒分组不发生改变

(1)第一步对数据进行规约化的意义在于方便后面余弦相似度的计算. 余弦相似度表示两个向量间夹角的大小, 其计算公式为:

$$Similar\ (\vec{a},\vec{b}) = \vec{a} * \vec{b} \Big/ \left|\vec{a}\right| * \left|\vec{b}\right|$$

计算结果为 1, 表示两向量完全相同, 若结果为–1, 表示两向量完全相反.

规约化的作用是将所有颗粒向量的模长规约为 1, 则余弦相似度公式可简化为:

$$Similar(\vec{a},\vec{b}) = \vec{a} * \vec{b}$$

(2)第二步的操作是假设分组初始中心. 因为对于聚类的每个分组, 都有一个代表该分组的中心. 在聚类开始之前, 并没有形成分组, 所以假设每个颗粒属于一个大的分组, 而假设第一个输入颗粒为初始聚类

的中心. 由于在后面的步骤中要对分组分裂以及对中心进行调整, 所以这两个假设不会对分组结果造成很大影响.

(3)第三步是该算法的核心步骤, 其又分为了 3 个步骤:

①计算输入颗粒与所有中心的余弦相似度, 并得到最大值(即找到与该颗粒最相似的分组).

②若最大余弦相似度大于阈值(该系统经试验得到最佳聚类阈值为 0.65), 则将该颗粒加入该分组, 并将自身颗粒特征加入中心, 对分组中心进行调整, 其公式为:

$$\overrightarrow{\theta_{\text{center}}} = \overrightarrow{\theta_{center}} + \lambda * \overrightarrow{\theta^{(i)}}$$

其中为 θ_{center} 中心向量, λ 为调整参数, 在该系统中这个参数为 0.05, $\theta^{(i)}$ 为新加入该组的颗粒向量.

③若最大余弦相似度小于阈值, 则该颗粒自身形成一个单独的分组, 并以自身作为新分组的中心.

(4)聚类分析结果:

分类的分组信息存储在数据库中, 如图 5.

GROUPID	MASSID
GTM2014103190813000	MASS-No_1-735585652407407360-00102
GTM2014103190813000	MASS-No_1-735585655486111100-01253
GTM2014103190813000	MASS-No_1-735585655983796220-01464
GTM2014103190813000	MASS-No_1-735585656608796160-01697

图 5　颗粒分组信息

GROUPID 表示分组编号, MASSID 表示颗粒编号. 具有统一分组编号的颗粒被聚类到同一组中.

聚类分析的结果可通过时间及采集地点查询. 图 6 表示输入所查询时间地点的界面:

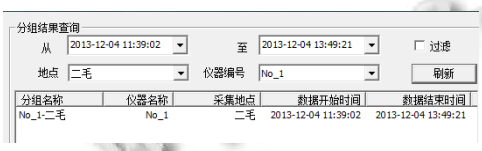


图 6　颗粒分组结果查询

该系统可以查询聚类后每个分组中所含有的颗粒数目, 如图 7 所示.

分组名称	颗粒数量
No_1-二毛{1}	669
No_1-二毛{2}	639
No_1-二毛{3}	517
No_1-二毛{4}	510

图 7　分组所含颗粒数查询

3.3 质谱数据自动命名

自动分类命名的功能是对于聚类分组进行成分的命名. 其核心算法使用数据挖掘重点逻辑回归模型. 自动命名的过程如下

自动分类命名算法

1、对每个聚类分组重新计算中心.

2、在所有中心随机抽取 20%作为训练样本, 其他分组作为测试样本.

3、使用训练样本训练各个颗粒种类的模型参数.

4、使用分类模型对测试样本分类.

5、人工验证分类模型性能.

(1)在完成聚类之后, 每个分组需要选取一个能够代表该分组的中心. 聚类自然形成的中心不一定处在最中心位置, 所以设定所有颗粒向量的平均值作为分组中心. 命名就是对分组中心进行的. 因为同一分组中颗粒相似度很大, 所以该组所含的颗粒都与中心命名一致.

(2)在分类算法中, 样本被分为训练样本和测试样本. 训练样本为分类训练模型参数, 测试样本测试分类的性能. 经试验, 该系统选取 20%的样本作为测试样本既可满足正确率的要求, 又可使人工工作量不至于过大.

(3)对于模型训练使用逻辑回归算法. 由于需要分类七种物质, 所以需要训练七个模型. 每个训练样本对七个模型都会产生作用, 其执行分为以下两个步骤:

①每种类型的模型都初始化为第一判定为该种类的样本. 如第 i 个样本是第一个被判定为 HM(重金属)的样本, 则 HM 的模型参数向量初始化为第 i 个样本, 即:

$$\overrightarrow{\theta_{HM}} = \overrightarrow{\theta^{(\mathrm{i})}}$$

②当每一个训练样本输入时, 根据人工判定的类别, 对每个类型的模型进行更新. 更新公式为:

$$\vec{\theta} = \vec{\theta} + \alpha\,(\overrightarrow{y^{(j)}} - \overrightarrow{\theta^{\mathrm{T}}}\overrightarrow{x^{(j)}})\overrightarrow{x^{(j)}}$$

其中 θ 代表模型参数; α 表示更新步长, 该系统中取值 0.05; $y^{(j)}$ 代表判别的类型, 可以取值 0 或 1, 如 j 颗粒的类型为 HM,则对于 HM 模型来说 $y^{(j)}=1$, 而对于 EC(单体碳)模型 $y^{(j)}=0$, 即每有　一个训练样本进入, 七个模型需要同时更新; $x^{(j)}$ 为输入样本向量.

(4)使用第三步训练出的模型对测试样本分类，分类的步骤为:

①计算测试样本与每个模型的余弦相似度.

②将样本命名为使得余弦相似度最大的类别.

(5)人工根据分组的颗粒平均图谱验证测试样本分类结果的正确率，若发现分类错误的样本，可手动修改，修改后模型自动执行第三步更新算法，对模型进行修正.

(6)自动命名结果

自动命名直接使用于聚类分组中，其结果在系统中显示，并且可以手动进行更改，如图 8 所示.

分组组名	所属类名
No_1-二毛{1}	NaK
No_1-二毛{2}	OC
No_1-二毛{3}	OC
No_1-二毛{4}	ECOC

图 8 分组命名结果

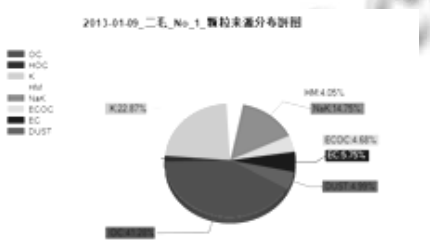
对于分组命名之后，分组中的所有颗粒都被命名，并将命名结果添加到数据库，结果如图 9 所示.

GROUPID	GROUPNAME	MASSCOUNT	NAME
GTM20141031908130000	No_1-二毛	21	EC
GTM20141031908130001	No_1-二毛	3	HOC
GTM20141031908130002	No_1-二毛	16	EC
GTM20141031908130003	No_1-二毛	4	K

图 9 颗粒命名结果

3.4 颗粒物成分统计

在进行完自动分类命名之后，所有的颗粒都已被命名，即成分已经确定. 为了对于环境状况进行评估，需要对大气颗粒中的成分进行统计. 图 10 为颗粒物成分分布的饼图.



10 颗粒成分分布饼图

图 11 为颗粒物成分随质谱仪中记录时间变化的堆叠图，可以看出颗粒物成分随时间的变化.

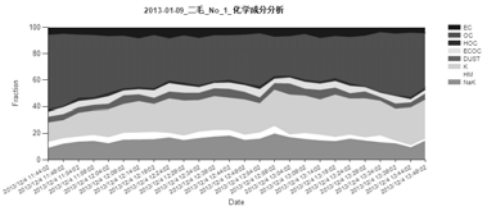


图 11 颗粒物成分随时间变化堆叠

4 结束语

本系统设计的主要目的是为了对于大气中的气溶胶颗粒进行实时在线的自动成分分析，使用了共振神经网络聚类算法和自动命名分类算法.

共振神经网络聚类算法有效的降低了数据规模，在环境污染较严重的情况下(质谱仪每天收集的颗粒超过 20 万)，分组数也不会超过 1000 组. 而自动命名算法在训练样本为 20%的情况下，正确率为 80%以上，完成了系统的预期目标.

参考文献

1 郭婧,华蕾,荆红卫.大气颗粒物的源成分谱研究现状综述.环境监控与预警,2011,3(6):28–32.

2 张莉.基于单颗粒气溶胶质谱信息的分类方法研究及其应用[学位论文].上海:上海大学,2013,5:1–87.

3 吉祥.数据挖掘技术在环境信息分析与预测中的应用研究[学位论文].苏州:苏州大学,2012:0–80.

4 王惠中,彭安群.数据挖掘研究现状及发展趋势.工矿自动化,2011,37(2):29–32.

5 石灵芝,邓启红,路婵,刘蔚巍.基于 BP 人工神经网络的大气颗粒物 PM10 质量浓度预测.中南大学学报(自然科学版),2012,43(5):1969–1974.

6 严悦.基于 ART 神经网络案例匹配的轨道交通智能数据诊断技术研究[学位论文].南京:南京理工大学,2013,8:1–62.

7 钱峰.国内数据挖掘工具研究综述.情报杂志,2008,27(10):11–13.

8 Yin YF, Gong GD, Han L. Air-combat behavior data mining based on truncation method. Journal of Systems Engineering and Electronics, 2010, 21(5): 827–832.