中国工程院院士时空数据分析及可视化

The Data Analysis of Academician of Chinese Academy of Engineering

专业方向：软件工程(人工智能方向)　　课程：人工智能高性能计算

小组成员：ZF1721106 曾凌芸　ZF1721115 崔程　ZF1721135 耿淼
ZF1721319 王少康　ZF1721350 杨航　ZF1721407 尹成浩

【小组成员】

ZF1721106 曾凌芸
ZF1721115 崔程
ZF1721135 耿淼
ZF1721319 王少康
ZF1721350 杨航
ZF1721407 尹成浩

【分工】

ZF1721106 曾凌芸：人员分工、数据整理、数据分析、PPT制作
ZF1721115 崔程：数据收集、数据整理、PPT制作
ZF1721135 耿淼：数据整理、系统开发、PPT制作
ZF1721319 王少康：特征提取、数据整理、PPT制作
ZF1721350 杨航：数据整理、数据分析、PPT制作
ZF1721407 尹成浩：数据整理、结果分析、PPT制作

01

Data Collection

数据收集

02

Feature Extraction

特征提取

03

System Development

系统开发

04

Data Analysis

数据分析

05

Valuable results

分析结果

Data Collection
数据收集

北京航空航天大學
BEIHANG UNIVERSITY

数据收集

scrapy框架+中文分词技术

1.查找数据源

2.分析网页结构

3.利用scrapy框架爬取数据

4.利用中文分词技术对数据做简单处理

数据收集

效果：

图1-院士数据源

图2-网页源代码

首先寻找工程院院士数据源，然后分析其网页源代码，以便爬取数据

数据收集

效果：

```python
# -*- coding: utf-8 -*-
import scrapy
from scrapy.linkextractors import LinkExtractor
from scrapy.spiders import CrawlSpider, Rule
from yuanshimingdangongcheng.items import YuanshimingdangongchengItem

class GongchengyuanshiSpider(CrawlSpider):
    name = 'gongchengyuanshi'
    #allowed_domains = ['http://www.cae.cn']
    start_urls = ['http://www.cae.cn/cae/sites/main/nav_qtysmd.jsp?ColumnID=48']

    contentlink = LinkExtractor(allow=r'cae/sites/main/jump')

    rules = (
        Rule(contentlink, callback="parse_item"),)


    def parse_item(self, response):

        item = YuanshimingdangongchengItem()


        item['url'] = response.url

        item['information'] = response.xpath('//div[@class="intro"]').xpath('string(.)').extract()[0].replace('\r','').replace(

        item['name'] = response.xpath('//div[@class="right_md_name"]/text()').extract()[0]

        yield item
```

图3-爬虫代码（部分)

北京航空航天大學 BEIHANG UNIVERSITY

数据收集

效果：

图4-爬取到的数据

北京航空航天大學 BEIHANG UNIVERSITY

数据收集

效果：

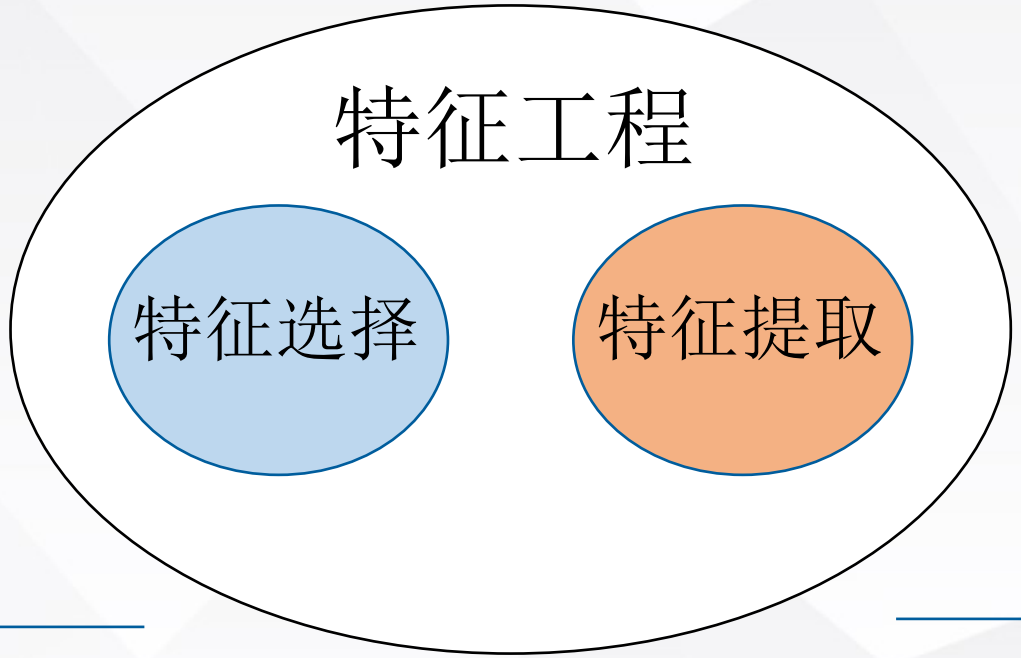
黄瑞松	中国航天科工集团公司、中国	huangruisong.jpg	院士	黄瑞松（1938.07.18- ）飞航技术专家。生于江苏省宜兴市
侯晓	中国航天科技集团第四研究院	20160504155634390664623.jpg	院士	侯晓（1963.10.8- ）航天固体火箭发动机专家。陕西岐山县
郭东明	大连理工大学	20120412145125370901075.jpg	院士	郭东明（1959.04.30- ）机械制造及自动化专家。河南省温
何琳	海军工程大学舰船振动与噪声	20180125160026601216710.jpg	院士	何琳（1957.11.11-），潜艇降噪技术专家。四川省西充县人
黄庆学	太原理工大学、重型机械教育	20180125160151757305267.jpg	院士	黄庆学（1960.12.05-），轧钢机械设计专家。吉林省舒兰市
黄先祥	中国人民解放军火箭军工程大	huangxianxiang.jpg	院士	黄先祥（1940.04.20- ）导弹发射与运用技术专家。江苏如
金东寒	中国船舶重工集团公司第七一	jindonghan.jpg	院士	金东寒（1961.01.11- ）动力机械工程专家。生于黑龙江省
李德群	华中科技大学、曾任华中科技	20160504155634419756634.jpg	院士	李德群（1945.8.7- ）材料成形专家。江苏省泰县人。1968
蒋庄德	西安交通大学	20140416142443297658264.jpg	院士	蒋庄德（1955.08.04- ）机械制造及自动
李骏	中国第一汽车集团公司	20140416142527729925232.jpg	院士	李骏（1958.03.24- ）汽车发动机专家。吉林省长春市人。
李椿萱	北京航空航天大学	lichunxuan.jpg	院士	李椿萱（1939.11.09- ）空气动力学、航
李培根	华中科技大学	lipeigen.jpg	院士	李培根（1948.12.27- ）机械制造及自动
李钊	西安工程兵工程学院	lizhao.jpg	院士	李钊（1940.02.03- ）地雷爆破专家。河北省无极县人。1964年毕业
李魁武	中国兵器工业集团第二〇二研	20160504155634325973892.jpg	院士	李魁武（1943.09.25- ）火炮自动武器及
林忠钦	上海交通大学	20120412145205633842542.jpg	院士	林忠钦（1957.12.06- ）机械工程专家。浙江镇海人。1982
龙乐豪	中国运载火箭技术研究院	longlehao.jpg	院士	龙乐豪（1938.7.4- ）武汉市人，研究员，火箭与航天技术
刘永才	中国航天科工集团公司、中国	LiuYongcai.jpg	院士	刘永才（1942.12.21- ）飞航技术专家。吉林省长春市人。
刘人怀	暨南大学、中国振动工程学会	liurenhuai.jpg	院士	刘人怀（1940.07.20- ）板壳结构分析与应用专家。四川省
刘怡昕	南京炮兵学院	liuyixi.jpg	院士	刘怡昕（1941.03.29- ）武器系统与运用工程专家。出生于
卢秉恒	西安交通大学机械工程学院	lubinheng.jpg	院士	卢秉恒（1945.02.05- ）机械工程专家。安徽省亳州市人。
孙逢春	电动车辆国家工程实验室	20180125160122836175843.jpg	院士	孙逢春（1958.06.06- ）车辆电动化专家。湖南省常德市

图5-分词结果展示

Feature Extraction
特征提取

特征提取

关系图



Feature Selection

在原始特征上进行
排序和选择

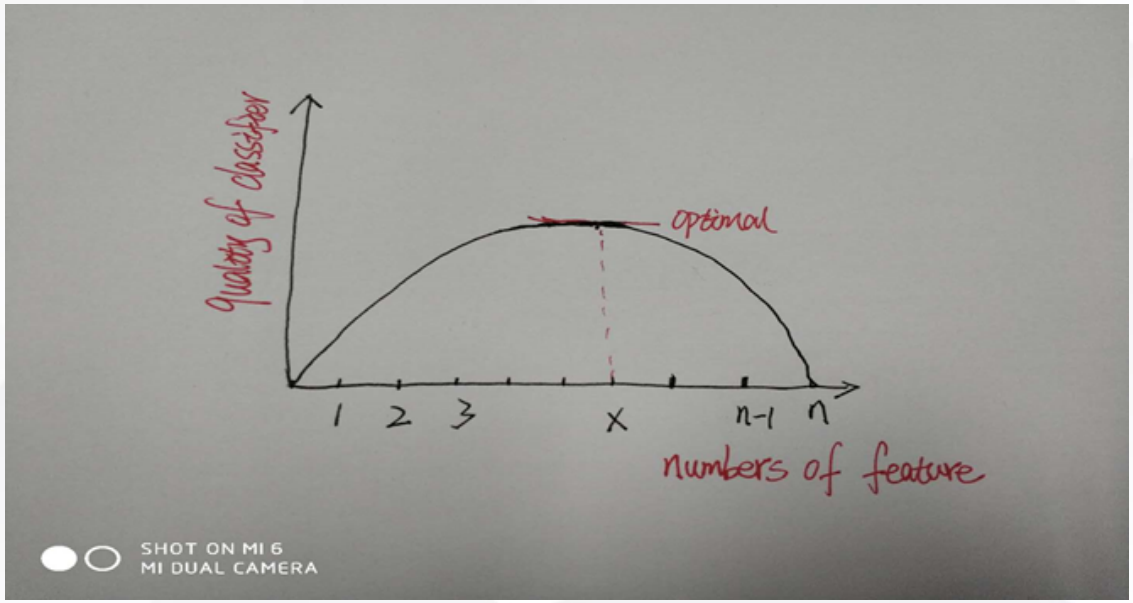
Feature Extraction

在原有特征基础上创造
凝练出一些新的特征

特征提取

特征选择与分类器性能关系

特征数量与分类性能



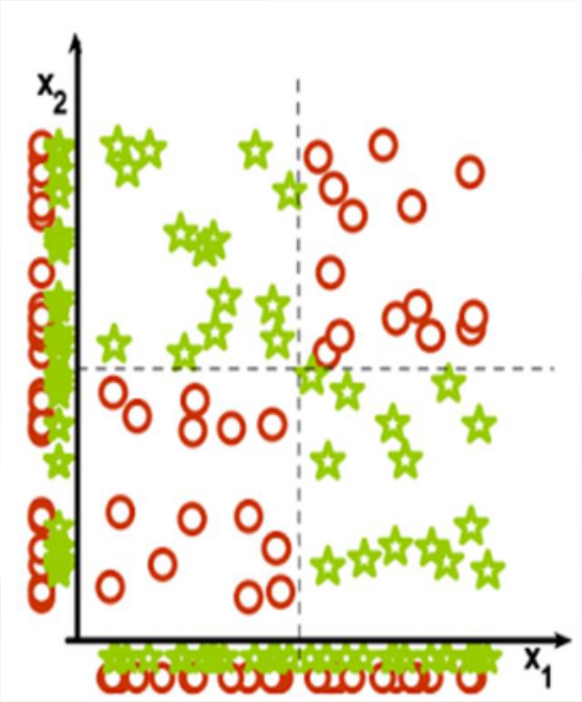
一般说来，当固定一个分类器的话， 所选择的特征数量和分类器的效果之间会有一个曲线，在某个x（1<=x<=n）的地方， 会达到最优。

特征提取

特征数量很重要

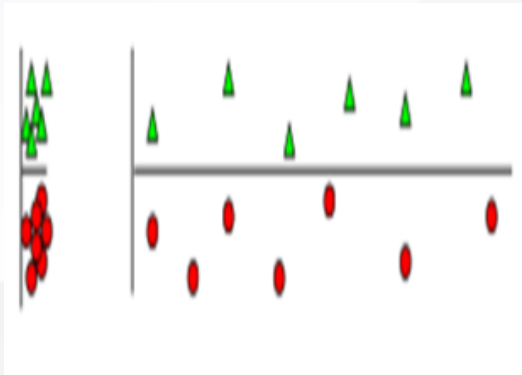
特征少

特征少了会导致无法区分的情况发生。 如右图1所示，仅仅依赖x1或者x2特征，都无法区分这两类数据， 所以当特征数量过小，很可能导致数据重叠。进而，所有分类器都会失效。



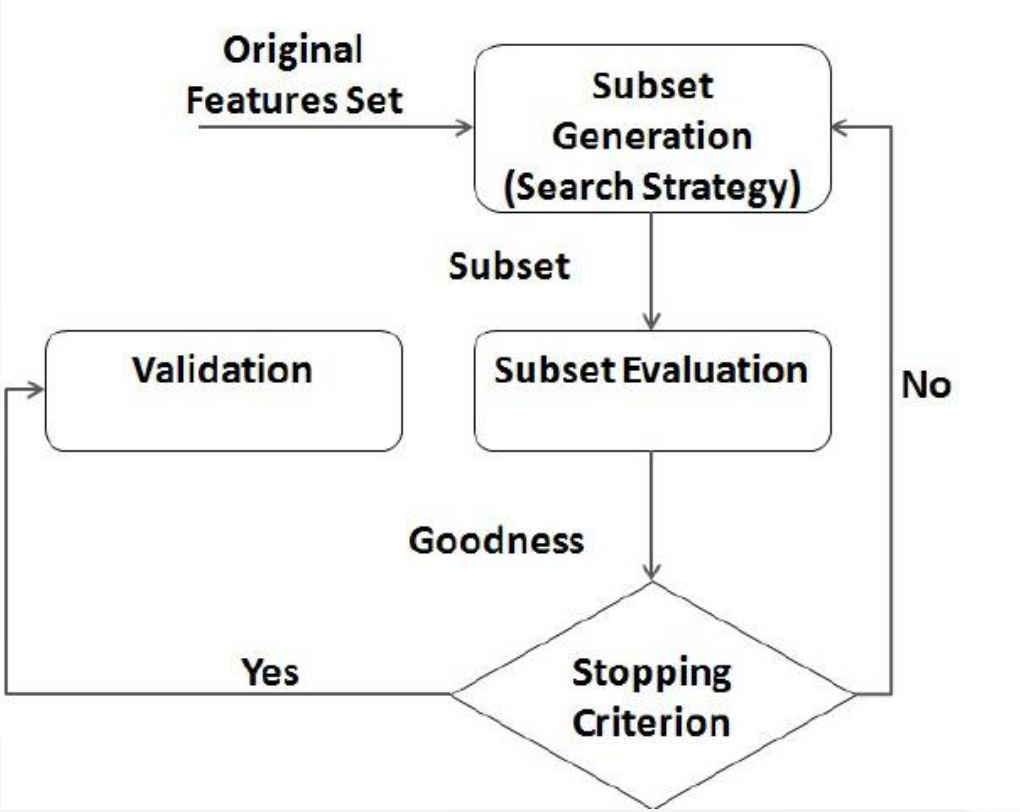
特征多

如右图2所示，根据纵轴来判断就可以容易的区分两类， 但是因为引入了横轴的特征， 使得同类数据在空间中距离变远，变稀疏了。 进而使得很多分类器失效。



特征提取

有用特征选择和提取



大数据时代的信息数据具有海量性、多样性、高速性和易变性，在带来数据信息的同时也带来了维度灾难，特征选择和提取就是去除嘈杂的或冗余的特征来减少数量的初始特征，选择一个子集来保留足够的有用信息来获得令人满意的结果。

System Development
系统开发

北京航空航天大學
BEIHANGUNIVERSITY

系统开发

- 开发工具：Visio Studio 2010 / Sql Server 2008R
- 开发语言：C# / sql / js / jquery / html / ajax
- 前后台通信：webservice / ajax
- 可视化呈现插件：highchart / map

系统开发

- C#是严格面向对象的编程语言，具有简单、易学的特点，可快速进行开发迭代；
- 使用sql语言进行数据库交互，可满足较大数据量的访问、读写操作；
- 前后台交互数据使用ajax / webservice，降低前后台通信数据量，提高用户体验度；
- 使用hightchart / map可视化工具，使得统计结果更加直观易懂；

系统开发

- C# 经典案例

```
37    public void GridBind(int pageindex, int pagesize) // 实现分页功能
38    {
39        pageindex = Convert.ToInt32(Session["pageindex"].ToString());
40        DataTable dt;
41        ArrayList objArray = new ArrayList();
42        objArray.Add(obDbType.GetParamBigint("PageNo", pageindex));
43        objArray.Add(obDbType.GetParamBigint("PageSize", pagesize));
44        dt = SqlHelper.GetTable("prGetT_professorMessagePage", objArray);
45        this.Gvlist.DataSource = dt;
46        this.Gvlist.DataBind();
47        if (dt.Rows.Count > 0)
48        {
49            this.AspNetPager1.RecordCount = int.Parse(dt.Rows[0]["total"].ToString());//总页数
50        }
51        this.AspNetPager1.PageSize = pagesize;//页面上显示的内容条数
52    }
```

系统开发

- webservice经典案例

```
26      [WebMethod]
27      [ScriptMethod(ResponseFormat = ResponseFormat.Json, UseHttpGet = false)]
28      public ArrayList getProfessorsStatic()//全部显示
29      {
30          ArrayList objs = new ArrayList();
31          string select = "SELECT *FROM (SELECT t.birthday,COUNT(t.birthday)AS number FROM (SELECT datepart(yy,B
    .T_professorMessage WHERE Birthday<'2000') AS t GROUP BY t.birthday) AS result ORDER BY result.birthday ASC";
32          DataTable dt = SqlHelper.GetTable(select);
33          foreach (DataRow dr in dt.Rows)
34          {
35              objs.Add(dr["birthday"] + "," + dr["number"]);
36          }
37          return objs;
38      }
```

系统开发

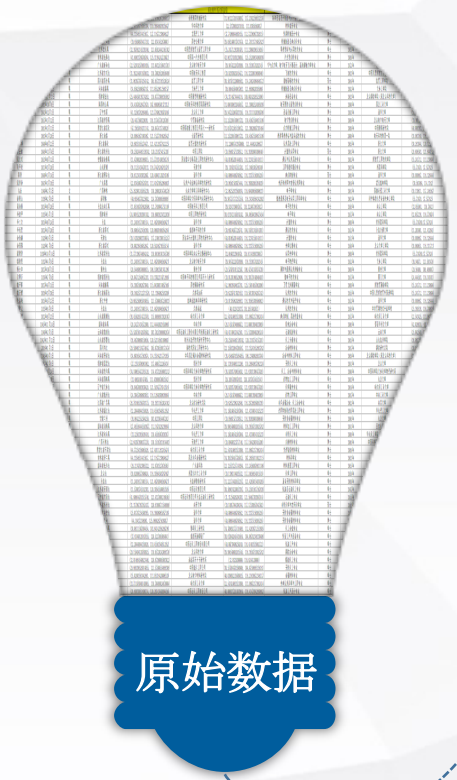
- js / jQuery经典案例

```
56    <script type="text/javascript">
57        function begin_show() {
58            var jsonDate = JsonNoParameter("webServices/ShowStaticServices.asmx/getProfessorsStatic");
59            var option = $("#id_select option:selected").val();
60            if (option == "columns") {
61                show_columns(jsonDate);
62            }
63            else {
64                show_pie(jsonDate);
65            }
66        }
67        function show_pie(jsonDate) {
68            var result = [];
69            var sum = 0;
70            for (var i = 0; i < 10; i++) {
71                result.push(i);
72            }
73            $.each(jsonDate.d, function (index, element) {
74                d = element.split(",");
75                var index = parseInt((d[0] - 1900) / 10);
76                result[index] += parseInt(d[1]);
77                sum += parseInt(d[1]);
78            });
```

Data Analysis
数据分析

北京航空航天大學
BEIHANG UNIVERSITY

数据分析



940x15

- 各实体对应的属性是一维的，不“鲜活”；
- 信息繁杂，潜在有价值的信息待挖掘；
- 数据属性不同质，需要采用一定的策略组合，进行分析和预测

北京航空航天大學
BEIHANG UNIVERSITY

数据分析

- 以地球为对象、基于统一时空基准，活动于时空中与位置直接或间接相关联的大数据

大数据

- 关联着过去、现在和将来，关联着世界万物的数字化、网络化和智能化，绝大部分数据都是时空大数据

- 现实地理世界空间结构与空间关系各要素的数量、质量特征及其随时间变化而变化的数据的总和

地理时空数据

- 包括基础地理信息数据、公共专题数据、智能感知实时数据和空间规划数据

北京航空航天大學
BEIHANG UNIVERSITY

数据分析

- 时变、空变、动态、多维演化。基于对象、过程、事件的时空变化是可度量的，其变化过程可作为事件来描述，通过对象、过程与事件的关联影射，可以建立时空大数据的动态关联模型

- 尺度特性，针对不同尺度的时空大数据的时空演化特点，可实现对象、过程、事件关联关系的尺度转化与重建，进而实现时空大数据的多尺度关联分析。

- 有多类型、多尺度、多维、动态关联特点，对关联约束可进行面向任务的分类分级，建立面向任务的关联约束选择、重构与更新机制，根据关联约束之间的相关性，可建立面向任务的关联约束启发式生成方法。

- 时间和空间两个维度，实时地抽取阶段性行为特征，以及参考时空关联约束建立态势模型，实时地觉察、理解和预测导致某特定阶段行为发生的态势。

北京航空航天大學
BEIHANG UNIVERSITY

数据分析

时空数据模型

基于状态与事件

状态和事件是时态地理信息系统的一对基本概念，根据这对概念可以对时空数据模型中时空对象建立时空拓扑关系，好的时空拓扑关系可以反应时空对象的时空演及内在的因果联系。

- 基于状态的模型
- 基于事件的模型
- 基于因果关系的模型

基于模型设计方法

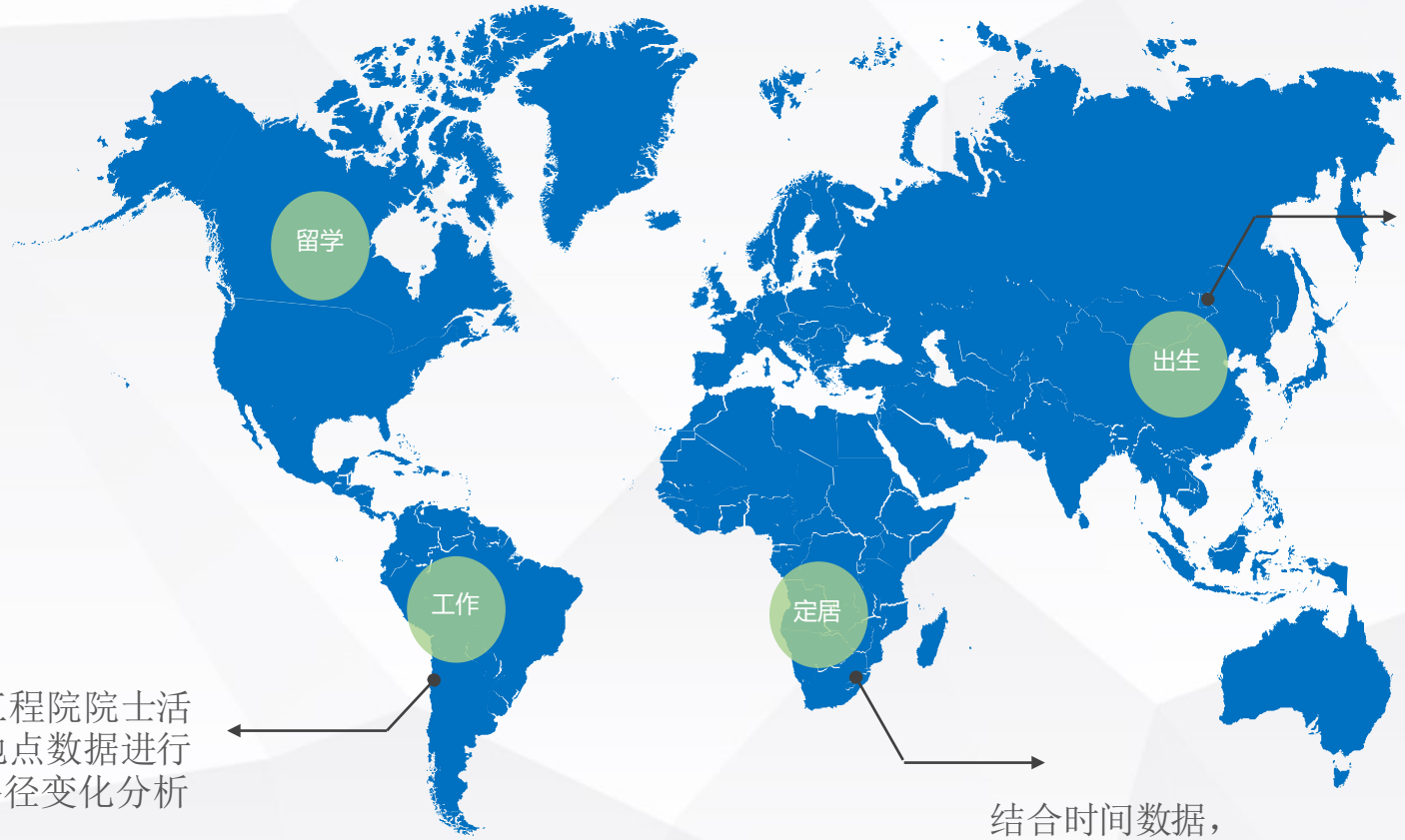
时空数据模型的研究集中了时态、空间及两者之间的唯一交互。目前，典型的时空数据模型设计方法有以下几种：

- 基于栅格的时空数据模型
- 基于矢量的时空数据模型
- 基于时间的时空数据模型
- 面向对象的时空数据模型

对比数据四：550页

北京航空航天大學
BEIHANG UNIVERSITY

数据分析



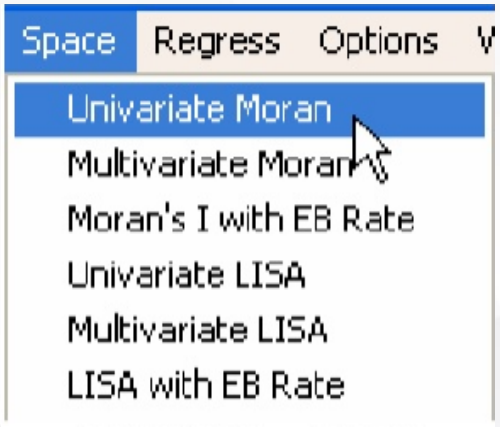
基于矢量的数据分析思想

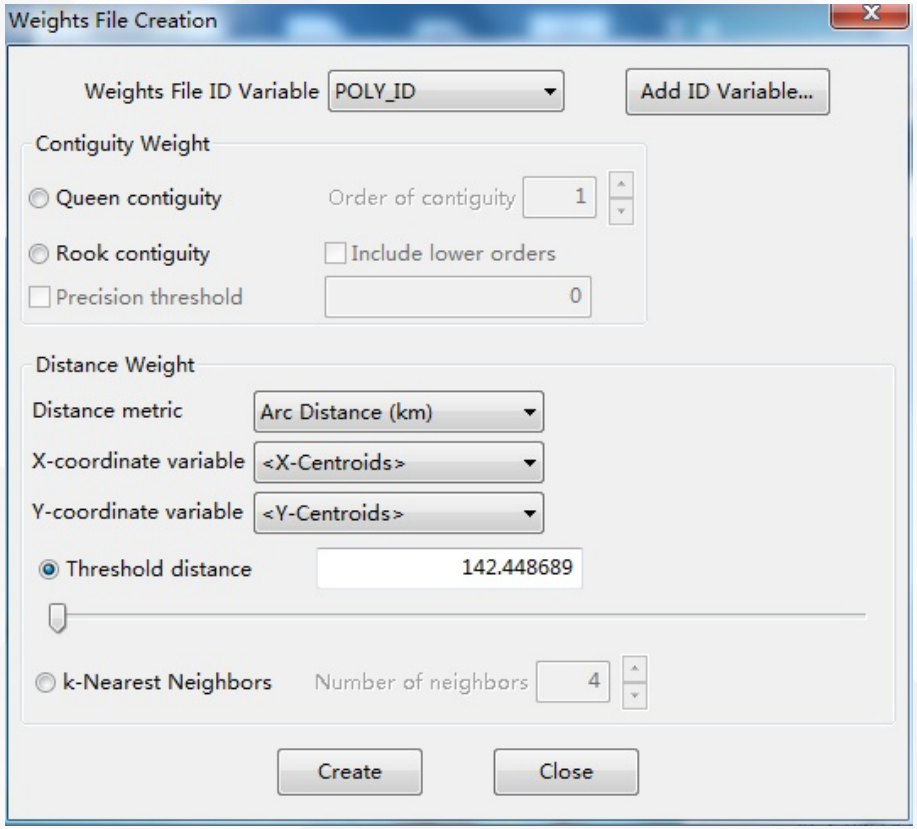
对工程院院士活动地点数据进行路径变化分析

结合时间数据，进行可视化综合展示

数据分析

分析工具：GeoDa



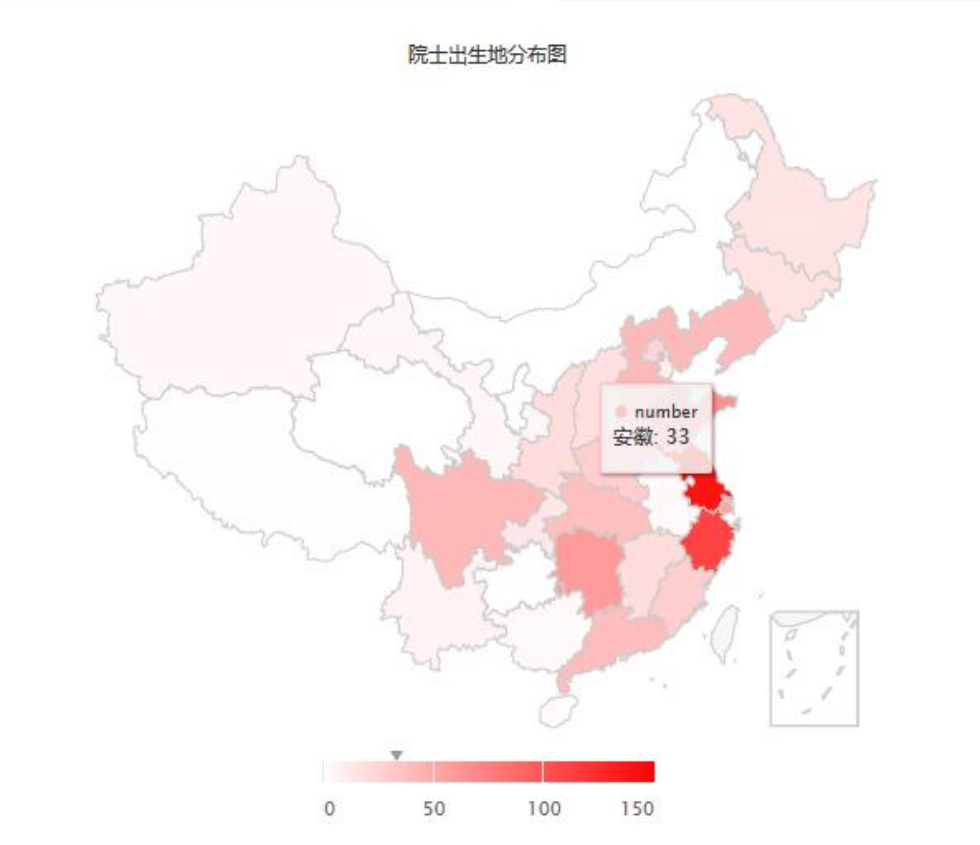


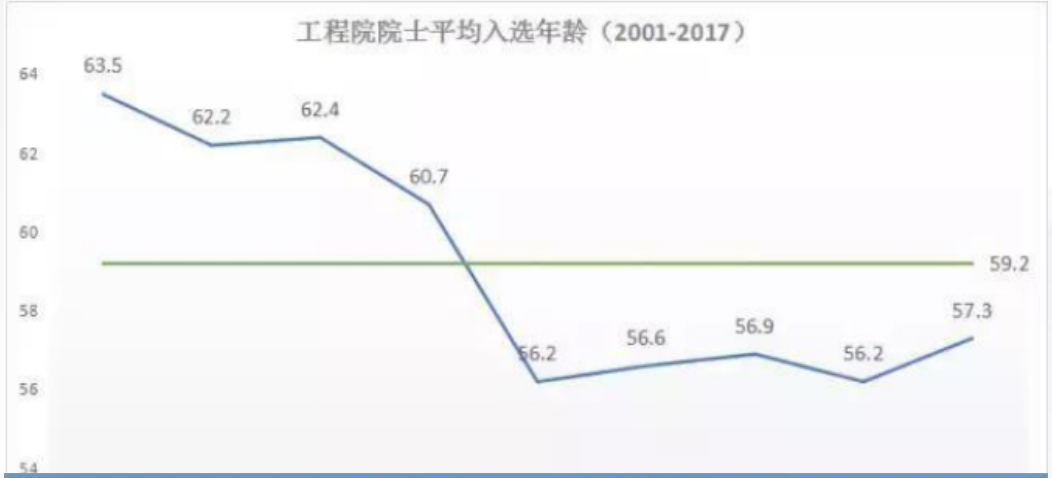
创建空间权重矩阵

北京航空航天大學
BEIHANG UNIVERSITY

数据分析

部分分析结果：





Valuable results
分析结果

北京航空航天大學 BEIHANG UNIVERSITY

分析结果

- 中国工程院院士出生地域分析

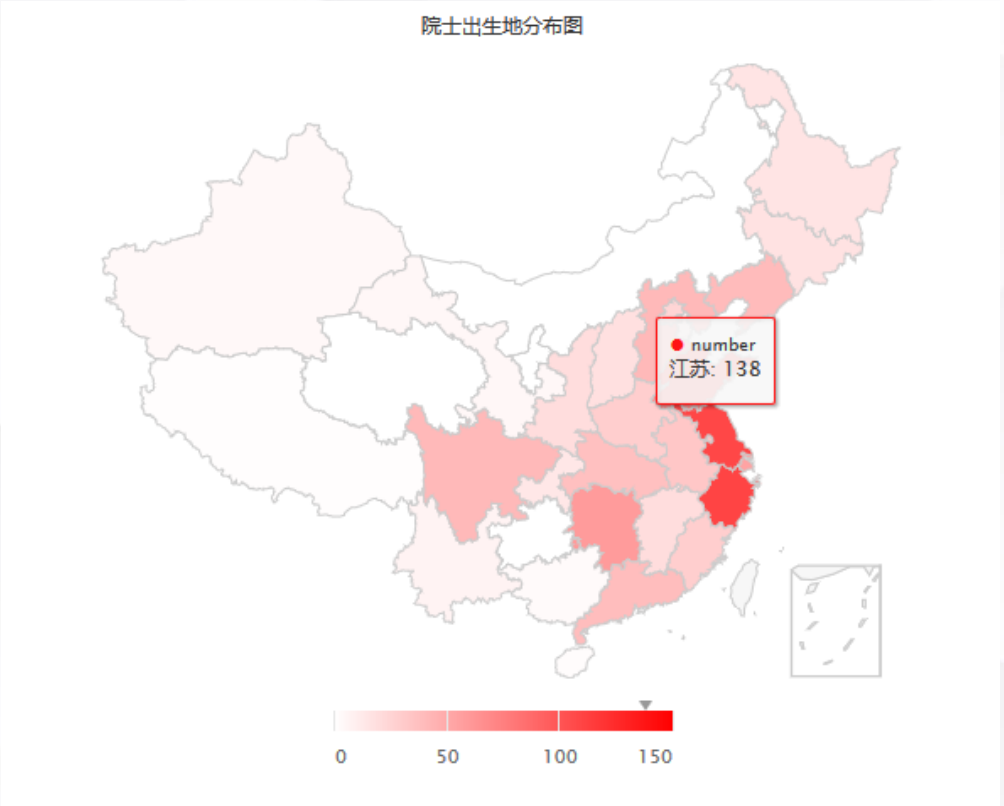
- 出生地域呈现特征如下：

1、地区之间分布不均衡。其中分布数量最多的位于江苏、浙江、上海、湖南、广东五省，而地处西北的甘肃、内蒙、贵州、西藏、新疆等省市地区出生的院士极少，我国工程院院士整体分布的态势为东南多而西北少。

2、集中于经济发达的省市。院士们主要集中分布在华东、华南、中南地区。经济文化较为发达、基础教育发展水平高的江苏、浙江和上海等省市，是院士的“高产区”。

- 分析结果：

院士们的出生地和工作地的空间分布极不均衡，院士们往往集中在经济文化发达、自然资源丰富的省份。在今后的科技人才的培养方向上我们应该立足于整个名族的繁荣进步，更多地强调地区平衡。



北京航空航天大學
BEIHANG UNIVERSITY

分析结果

- 中国工程院院士出生年份分析

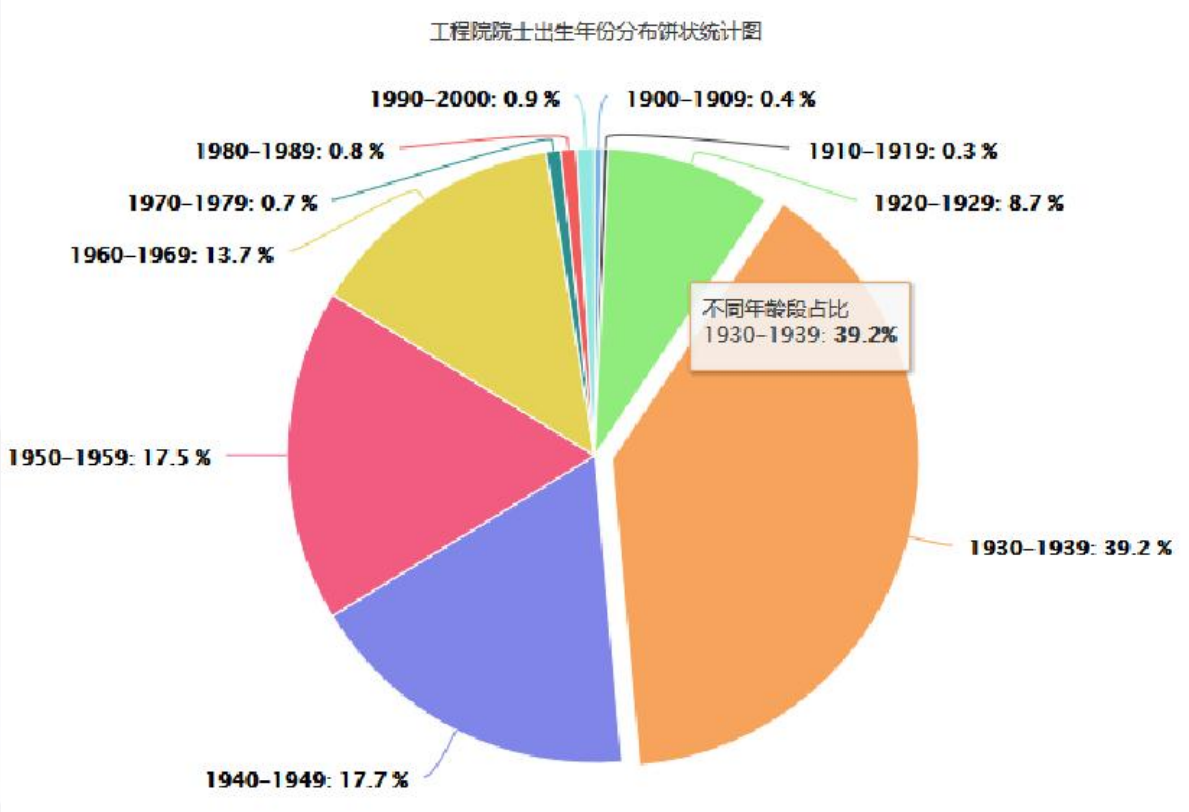
- 出生年份呈现特征如下：

1、1930-1939年出生的院士最多，占比39.2%，其次是1940-1949年，占比17.7%，1950-1959年出生院士占比17.5%。

2、1900-1919年和1970-2000年出生的院士最少，总共占比之和为3.1%。

- 分析结果：

我国工程院院士主要群体普遍达到了退休年龄或者已经退休，年轻群体的院士较少，由此可以推断，我国科研群体缺乏新鲜血液，缺乏创新力与创造力。



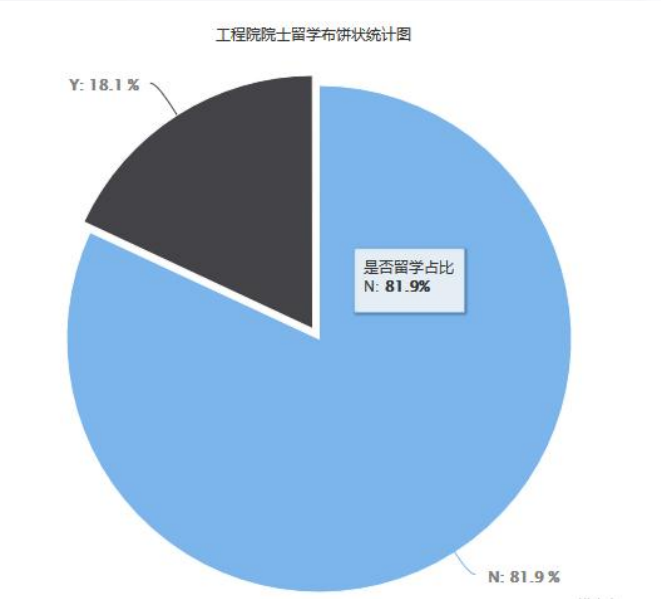
北京航空航天大學 BEIHANG UNIVERSITY

分析结果

- 中国工程院院士留学状况与最高学历分析

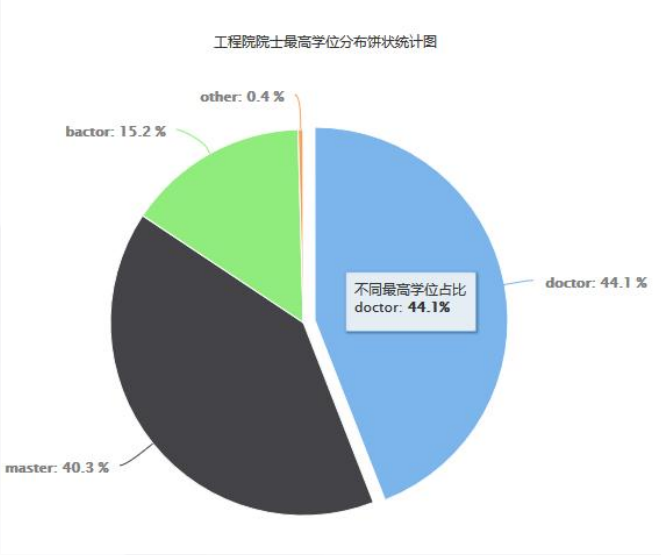
- 呈现特征：

1、我国工程院院士有过留学经历的占比18.1%。
2、我国工程院院士拥有博士学位占比最高，占44.1%，其次是硕士学位占40.3%。



- 分析结果：

目前我国工程院院士大部分没有出过留学经历，可以推断是受几十年前我国的国情与经济水平影响，我们国家没有条件让更多人出国深造。院士绝大部分具有硕士及以上学历，因此可以推断我国工程院院士选举条件较为苛刻，对学历要求比较高。



北京航空航天大學
BEIHANG UNIVERSITY

分析总结

针对以上所有数据分析，我们给出如下总结：

第一，中国工程院院士是中国科技界的精英，是中国科学技术现代化的开创者和发展的中流砥柱，我们应重视和保护这一精英群体。但我国工程院院士地域分布不均衡，各省份之间差别悬殊，这种人才分布不均衡的发展只会加剧地域发展的不平衡性，这对于我国人才的培养百害而无一益。

第二，留学教育对于我国科技人才发展具有重要意义。因为经济条件和政治因素的影响，使得我国工程院院士留学比率只有18.1%，但是现如今我国经济条件飞速发展，教育水平日渐完善，我们国家有更加充沛的实力让更多的学子去国外深造，学习国外最先进的技术，为日后报效祖国打下坚实的基础。

第三，我国工程院院士的年龄结构状态堪忧，通过数据分析，我们可以得知，现如今绝大部分院士的年龄均已达到退休年龄，年轻的院士数量占比太少，我国工程院院士整体而言缺乏新鲜血液，这对于我国科技的创新与发展是没有任何帮助的，我们必须更加清醒得认识到这一问题的严峻性，让更多富有年轻活力，创造力的年轻人加入到我国的科研工作之中。

THANK FOR YOUR WATCH

恳请老师批评指正