第三届泰迪杯全国大学生数据挖掘竞赛试题

说明：1、参赛选手可从下述试题中任选一题作答，并在论文报告中标明

2、论文等级会综合考虑论文质量和难度系数

试题一 基于电商平台家电设备的消费者需求及产品数据挖掘分析（难度系数：1.0）

试题来源：美的 Midea

背景：

随着互联网与移动互联网的快速发展，截止 2014 年 6 月，我国的网民规模达 6.32 亿，互联网普及率为 46.9%，2015 年中国网民的渗透率将接近 50%。2014 年天猫双十一的交易额达 571 亿，网上购物将成为人民生活的一部分。网民在电商平台上浏览和购物，产生了海量的数据，如何利用好这些碎片化、非结构化的数据，将直接影响到企业产品在电商平台上的发展，也是大数据在实际企业经营中的应用。对于用户在电商平台上留下的评论数据，运用文本分析方法，了解用户的需求、抱怨，购买原因以及产品的优点、缺点，对于改善家电设备产品及用户体验有着重要的意义。

据观研天下行业分析：近年来我国家电设备销量增长迅速，以电热水器为例，2011 年电热水器市场销量比 2010 年增长 2.29%，销售额增长 5.23%；2013 年热水器零售量达到 2842 万台，零售额达到 459 亿元，2014 年热水器整体规模向上，但增速较 2013 年有所回落，零售量达到 2985 万台，零售额达到 504 亿元。

需求：

1、 分析用户对于热水器/净水器产品的个性化需求；
2、 分析现有电商热水器/净水器的产品劣势（用户抱怨点）及产品优势（用户赞点）；
3、 分析各品牌的产品间的差异，进行差异化卖点提炼；
4、 分析用户购买的原因；
5、 对用户的购买行为进行分析挖掘（搜索关键字、购买时关注点、购买步骤、使用、评价）（此部分可选择来做）。

提示：

1、 在电商平台进行评论数据抓取（可用火车头采集器进行评论爬虫）；
2、 对评论数据进行预处理（处理掉水军及随意发表的评论数据）；
3、 可分品类进行细化分析（热水器：电热热水器、燃气热水器；净水器：净水机、纯水机）；
4、 对评论数据进行文本分析（好评、差评、中文分词、词频统计、情感分析、语义网络）；
5、 可利用百度指数、淘宝指数等互联网工具对热水器和净水器的消费人群及搜索关注点进行分析；
6、 建议在国内外相关文献的基础上尽量选择新技术手段进行挖掘，比如基于深度学习理论模型完成情感分析，参见文献：《基于深度学习的微博情感分析》、《基于深度学习的文本情感分类研究》等。

说明：

1、 **附件一_电热水器及净水器评论数据集.zip** 是电热水器和净水器的评论数据，参赛者也可去电商平台进行数据爬虫，爬取最新的评论数据（若自行爬取的数据，提交论文成果时请一并提交）；

2、 参赛者可以选择热水器或者净水器任一一个品类进行分析挖掘；

3、 参赛者可以从以上需求选择部分或所有主题进行建模分析，也可提出自己的分析主题。

试题二 基于数据挖掘技术的市财政收入分析预测模型（难度系数：0.8）

试题来源：泰迪智能科技 TipDM Intelligent Technology

背景：

在我国现行的分税制财政管理体制下，地方财政收入不仅是国家财政收入的重要组成部分，而且具有其相对独立的构成内容。地方财政收入是区域国民经济的综合反映，也是市场经济国家的政府进行宏观调控的基础。科学、合理地预测地方财政收人，对于克服年度地方预算收支规模确定的随意性和盲目性，正确处理地方财政与经济的相互关系具有十分重要的意义。

广州市作为广东省的省会，改革开放的前沿城市，交通便利，拥有中国大陆三大国际航空枢纽机场之一的广州白云国际机场和中国第三大港口、港口货物吞吐量居世界港口第五位的广州港。广州号称“千年商埠”，历史上一直是中国最重要的商业中心之一，商业网点多、行业齐全、辐射面广、信息灵、流通渠道通顺，拥有商业网点 10 万多个，为中国十大城市之冠。广州市在实现经济快速发展，地区生产总值飞跃的同时，也意味着财政收入的增收。2013 年，广州实现地区生产总值（GDP）15420.14 亿元，增长 11.6%。其中，第一产业增加值 228.87 亿元，增长 2.7 %；第二产业增加值 5227.38 亿元，增长 9.2%；第三产业增加值 9963.89 亿元，增长 13.3%。第一、二、三次产业增加值的比例为 1.48∶33.90∶64.62。三次产业对经济增长的贡献率分别为 0.4%、29.0%和 70.6%。广州地方公共财政预算收入 1141.79 亿元，增长 10.8%；如何做出下一年有效的财政收入预算，为下一年的政策提供指导依据，是一个具有重大意义的问题。

需求：

1、 梳理影响广州市财政收入关联指标的有关数据，分析、识别影响财政收入的关键影响因素；

2、 结合需求 1 的因素分析，利用相关的数据挖掘技术对广州市 2015 年的财政总收入及各个类别收入进行预测；

3、 结合社会经济发展和广州市近几年的财政收入及支出等情况，从财政收入和支出预算的角度，向广州市财政局提出几点建议。

提示：

1、 可在广州市统计信息网（http://www.gzstats.gov.cn/）下载相关数据；

2、 在税收方面，可进行细化分析，如增值税、营业税、企业所得税、个人所得税等；

3、 在向广州市财政局提建议时，考虑经济因素和非经济因素；

4、 建议查找多方面的数据，进行综合分析。

说明：

1、 **附件二_相关资料.zip** 为财政收入影响因素的参考资料，参赛者也可自行查找相关资料。

试题三 城市供水处理混凝投药过程的建模与控制（难度系数：0.9）

试题来源：GDH 广东粤港供水有限公司 GUANGDONG YUE GANG WATER SUPPLY CO.,LTD.

背景：

水是生命的源泉，是人类生活不可缺少的成分，然而随着工业发展迅速，人类活动范围的快速扩大，水资源受到的污染日益严重。因此，怎么样有效地对水进行净化处理，成为了当今国内外学者研究的热点问题。

对水进行净化处理要经过混合、絮凝、沉淀、过滤和消毒五个阶段，絮凝沉淀是水处理的初始环节，是悬浮颗粒、胶体等杂质处理的必需工艺。影响絮凝效果的因素很多，包括原水流量、原水浊度、原水 pH 值、原水温度、混凝剂投加量和原水中藻类等等。投药控制就是综合考虑这些因素进行混凝剂最少最经济投加，而达到最优的絮凝效果。浊度为水的清亮程度，是水质指标的重要参数，单位为 NTU。混凝就是用混凝剂把水中胶体粒子以及微小悬浮物的聚集过程，是凝聚和絮凝的总称，凝聚是胶体失去稳定性的过程，絮凝是脱稳胶体相互聚集，沉淀则是将混凝后的水中凝聚物实现下降、沉积，减少上层水中的凝聚物数量。通过混凝和沉淀就可以减少水中悬浮颗粒的数量和大小，也就能实现浊度降低，投药控制的目的是使沉淀池的出水浊度符合相关标准。此外由于混凝沉淀池是一个大容积对象，因此对于混凝剂投加与对应水絮凝沉淀后的浊度存在一段较长的时间差，造成控制滞后。图 1 展示了投药控制流程。

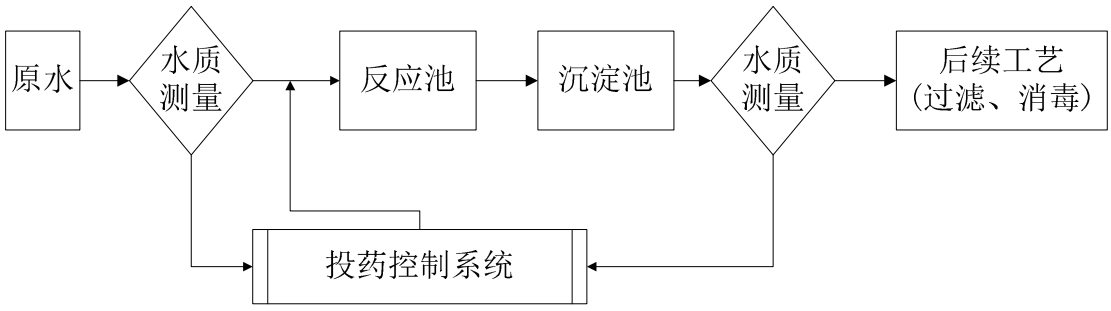

图 1 投药控制流程

水处理混凝投加过程是一个复杂的物理、化学反应过程，具有时滞和非线性特性。目前的混凝投药控制方法中总有一些不足之处，如烧杯实验法需要每天或每周进行频繁试验，耗时很多且对输出水质影响很大；流动电流法中的流动电流检测器在使用过程中会逐渐降低精度，且在高浊度水或某些污染较严重的水质和絮凝剂是有机阴离子高分子时不能适用；数学模型法因混凝过程复杂，难以建立高精度和高可靠性的过程模型导致控制不能适应控制情况

的变化，所以总的来说目前的投药控制方法都是难以适应水质的变化，鲁棒性较差、抗干扰能力较弱。

本案例的水厂在抽取原水后会进行化学预氧处理，达到除去微量有机污染、除藻、除臭味、控制氯化消毒副产物、氧化助凝和除去铁锰等目的。水厂选用混凝剂是 PAC，添加混凝剂后的水在反应池进行絮凝，流入 3 号和 4 号沉淀池，取 3 号和 4 号沉淀池出水浊度的平均值作为沉淀池出水浊度，沉淀池出水浊度的合格标准为不大于 1.10NTU。在历史数据中，存在药剂反应效果不好，沉淀池出水浊度不合格的数据。一般情况下，原水添加混凝剂反应到沉淀结束出水需要 70min 到 120min。

需求：

1、 根据历史原水水质数据、原水流量数据、混凝剂投加量和沉淀池出水浊度数据，求出原水添加混凝剂反应到沉淀结束出水需要的时间。
2、 考虑需求 1 结果的滞后性，根据历史原水水质数据、原水流量数据和混凝剂投加量数据，建立数学模型，求出最佳混凝剂投药量。
3、 考虑需求 1 结果的滞后性，考虑增加沉淀池浊度作为输入参数，结合历史原水水质数据、原水流量数据和混凝剂投加量数据，建立数学模型，求出最佳混凝剂投药量；
4、 通常而言，温度也是影响化学反映速度的一个重要因素。原数据中并未包含温度数据。请做出相应的尝试引入温度数据，并分析其对最佳投药量的影响。

提示：

1、 水处理过程的最终目标是通过对原水水质参数的分析，在线实时控制药剂的投加量，以适应原水水质的不断变化，使出水满足各项水质指标。即根据历史数据辨识建立进水流量、浊度、PH 值、加药量和沉淀池出水浊度之间的数学模型，实时确定最佳混凝剂投药量。

说明：

1、 **附件三_投药控制数据集.zip** 是某水厂投药控制系统实时采集的数据信息，数据均为瞬时测量值，取水量是原水的流速，供水量是出厂水的流速（沉淀池后还有部分工艺会造成水的损耗），取水量和供水量的单位是 m^3/h（立方米每小时），PAC 耗是混凝剂 PAC 的消耗，单位是 mg/L（1L 原水消耗 PAC 的量）；
2、 参赛者可以从以上需求选择部分或所有主题进行建模分析。