



泰迪智能科技

服务热线：40068-40020

顶尖数据挖掘教学实训平台 之案例库（TipDM-C10）

产 品 说 明 书

广州泰迪智能科技有限公司 版权所有

地址：广州市经济技术开发区科学城 232 号

网址：<http://www.tipdm.com>

邮箱：services@tipdm.com

热线：40068-40020

企业 QQ：40068-40020

邮编：510663

电话：（020）82039399

修订记录

日期	版本	作者	修改内容
2016-7-5	1.0	放榜松	初稿
2016-10-28	2.0	莫芳	修改案例库清单
2016-10-31	2.1	施兴	修订部分案例文字说明、格式及增加配套资源 ppt 及微课部分

目 录

1	案例库概述	4
2	案例构成.....	26
3	案例示例.....	27
3.1	案例介绍.....	27
3.1.1	案例与挖掘目标.....	27
3.1.2	分析方法与过程.....	28
3.1.3	上机实践.....	28
3.1.4	拓展思考.....	29
3.2	配套资源.....	29
3.2.1	配套程序.....	30
3.2.2	配套数据.....	30

1 案例库概述

顶尖数据挖掘教学实训平台之案例库（TipDM-C10）是泰迪科技在数据挖掘领域探索 10 余年和高校资深讲师联合经验总结之作，内容涵盖智能电网、移动通信、医疗健康、网络舆情、电子商务、金融保险、交通运输、信息安全、政务民生等诸多行业，特别适合有数据挖掘相关课程教学的高校、研究所和培训机构，也可作为个人学习数据挖掘技术的最佳素材。

数据挖掘案例主要包括为：

行业类别	案例名称	案例描述	主要知识点	应用平台
电力行业	电力窃漏电用户自动识别	通过采集电力计量自动化系统的电流、电压、功率因数等用电负荷数据，用电异常等终端报警信息以及电力营销系统提供的历史窃漏电用例，归纳出窃漏电用户的关键特征，构建出窃漏电用户的识别模型，最终通过监测用户用电数据，实现窃漏电用户的实时在线诊断。	1.分布分析 2.周期性分析 3.拉格朗日插值 4.CART 决策树 5.LM 人工神经网络 6.混淆矩阵 7.ROC 曲线	TipDM-TB TipDM-HB R 语言 Python MATLAB
	气象与输电线路缺陷关联分析	通过分析鸟害、接地装置受损、保护区安全隐患、部件发热异常、锈蚀损伤等缺陷与气象因子的关联关系，构建出气象因子与缺陷关联关系的预警模型。实现在极端气象条件下缺陷	1.分布分析 2.趋势分析 3.Pearson 相关系数 4.数据去重 5.牛顿插值	TipDM-TB R 语言 Python

	在高发区发生的概率并进行预警。	6.拉格朗日插值 7.因子关联分析	
电能计量 异常智能 诊断	通过构建电能计量异常评价指标体系及电能计量采集异常专家样本库和具有自适应和自学习特性的电能计量异常智能诊断模型,准确判断出计量装置损坏真因。	1.特征提取 2.C4.5 决策树 3.BP 神经网络 3.混淆矩阵 4. ROC 曲线	TipDM-TB R 语言 Python
用电客户 分群研究	企业信息化建设以及下一代智能化电网的全面建设,电力数据资源急剧增长并形成了一定规模。构建客户细分模型,将电网中的所有用电客户细分成五个不同的群体,分析模型得出的各个群体的特征,根据不同的群体,为电网提出差异出的营销建议。	1.Oracle SQL 查询提取数据 2.箱线图异常值检测 3.K-means 算法	TipDM-TB R 语言 Python
电力客户 满意度综合评分	供电企业的的基本任务是提供给客户电力商品与服务,满足客户需求的同时获得盈利。客户较高的满意度是企业长期奋斗的目标,需要从客户需求出发,制定相关策略,实现客户满意度评价。根据电力营销系统各模块原始数据,构建电力客户满意度评价指标体系及满意度综合评分模型。	1.分布分析 2.层次分析法 3.熵权法（客观权重） 4.混淆矩阵 5. ROC 曲线	TipDM-TB R 语言 Python

移动通信	基于基站定位数据的商圈分析	商圈划分的目的之一是为了研究潜在的顾客的分布以制定适宜的商业对策。从通信运营商提供的基站定位数据,利用基站小区的覆盖范围作为商圈区域的划分,归纳出商圈的人流特征和规律,识别出不同类别的商圈,实现营销指导。	1.特征提取 2.零-均值标准化 3.层次聚类法	TipDM-TB TipDM-HB R 语言 Python MATLAB
	基于大数据的运营商用户流失分析	随着大数据挖掘技术的不断发展和应用,移动运营商希望能借助数据挖掘技术识别哪些用户可能流失,什么时候会发生流失。而通过建立流失预测模型,分析用户的历史数据和当前数据,提取辅助决策的关键性数据,并从中发现隐藏关系和模式,进而预测未来可能发生的行为,就可以帮助移动运营商实现这些要求	1.特征分析及提取 2.欠抽样与过抽样 3.逻辑回归 4.支持向量机 4. ROC 曲线 6.召回率 7.F 值	TipDM-TB R 语言 Python
	金融理财的广告牌精准投放	从线上信息的浏览归纳出不同客户对金融理财的个体偏好特征,构建模型对客户进行分类,识别对金融理财感兴趣的重点和潜在客户群。进一步对重点和潜在客户群的时空轨迹进行定位,确定热门聚集地和聚集时	1.K-means 聚类分析 2.LOCF 插值法 3.轨迹挖掘	TipDM-TB R 语言 Python

		点,从而为金融理财产品广告牌的精准营销和投放提供理论依据。		
手机终端 用户变迁 预测		通信市场逐渐趋于饱和,想依赖发展新客户群带动运营商利益的时代已不复存在。而且,现今的客户在更换手机的频率和对手机的依赖度将越来越高,所以对客户换机倾向的预测势在必行。归纳出换机用户的关键属性与行为特征,构建换机用户的识别模型,并预测用户后一个月的换机情况。	1.分布分析 2.趋势分析 3.SMOTE 算法 4.决策树 5.神经网络 6.交叉验证	TipDM-TB R 语言 Python
电信业务 客户价值 评价实证 分析		对于电信运营商而言,任何时候都不可以松懈的任务就是如何扩大客户资源,如何保留已有的客户资源。基于客户属性与历史行为数据,从客户的当前价值及客户的潜在价值两个维度出发,构建客户价值评价指标体系;根据指标体系构建客户价值评价模型,识别不同价值等级的客户,并针对不同的群体为企业资源配置提供意见和建议。	1.层次分析法 2.主成分分析 3.K-means 聚类分析	TipDM-TB R 语言 Python

医疗健康	中医证型 关联规则 挖掘	借助三阴乳腺癌患者的病理信息,发现中医症状间的关联关系和诸多症状间的规律性,挖掘潜在证素,并且依据规则分析病因、预测病情发展以及为未来临床诊治提供有效借鉴。这样患者在治疗过程中,医生可以有效的减少西医以及化疗治疗的毒副作用,为后续治疗打下基础。	1.问卷调查 2.问卷结果数值化 3.数据 Kmeans 离散化 3.Apriori 算法	TipDM-TB R 语言 Python MATLAB
	脑中风发病环境因素分析及干预	对脑中风的发生环境因素进行分析,其目的是为了进行疾病的风险评估,对脑中风高危人群能够及时采取干预措施。根据病人基本信息,对发病人群进行统计描述。建立数学模型研究脑卒中发病率与气温、气压、相对湿度间的关系。查阅和搜集文献中有关脑卒中高危人群的重要特征和关键指标,对高危人群提出预警和干预的建议方案。	1.周期性 2.自相关性 3.多重共线性分析 4.过度离势检验 5.泊松回归 6.违背经典假设检验	TipDM-TB R 语言 Python

基于基因表达谱的肿瘤识别	基因表达谱能够提供丰富的信息进行正常和患有肿瘤两类样本的辨别,为医学诊断及抗癌药物研制便捷。选取一定标准,作为衡量某基因是否为“无关基因”的判断条件,然后对样本的基因表达谱进行筛选,剔除与肿瘤无关的基因。设计分类器提取肿瘤相关基因标签。将噪声干扰考虑到基因表达谱的分析中,分析通过引入噪声模型对基因标签的确定的影响。分析基因标签判断的准确性。	1.数据归一化 2.特征提取 3.巴氏距离 4.FSSM 算法 5.支持向量机 6.留一法交叉验证 7.小波阈值去噪法 8. LDA-GA 方法	TipDM-TB R 语言 Python
医保欺诈用户的自动识别	医疗保险欺诈,是指公民、法人或者其他组织在参加医疗保险、缴纳医疗保险费、享受医疗保险待遇过程中,故意捏造事实、弄虚作假、隐瞒真实情况等造成医疗保险基金损失的行为。根据医保欺诈数据,建立自动识别模型,识别医保欺诈记录。	1.比较法 2.Benford 定律 3.K-means 聚类 4.关联规则	TipDM-TB R 语言 Python

	市政人口 与医疗需 求预测	某市外来人口较多，人口密度过大，给交通、医疗、教育等社会公共服务带来了巨大的压力。根据 1979 年至 2010 年某人口数据，建立模型预测未来五年某市户籍、非户籍、常住人口数量；根据 2000 年、2005 年、2010 年三年的深圳市常住人口年龄结构，预测 2015 年某市常住人口年龄结构；收集数据，以病床需求量为参考，预测 2014 年某市医疗需求状况。	1.ARIMA 时间序列 2.Leslie 人口模型	TipDM-TB R 语言 Python
网络舆情	电商评论 数据情感 分析	在电商平台激烈竞争的大背景下，除了提高商品质量、压低商品价格外，其中非常重要的方式就是对消费者的文本评论数据包含的内在信息进行挖掘分析。而得到的这些信息，也会有利于对应商品的生产厂家自身竞争力的提升。分析某一热水器的用户情感倾向，从评论文本中挖掘出该热水器的优点与不足，提炼不同品牌热水器的卖点。	1.八爪鱼爬取数据 2.机械压缩去词 3.吉布斯抽样 4.LDA 主题分析	TipDM-TB TipDM-HB R 语言 Python MATLAB

基于网络招聘信息的热门职位分析与人才需求预测	对网络招聘信息进行分析研究,了解不同地区和行业的需求特点,分析热门地域和行业,挖掘兴起的数据类行业相应的人才需求现状及发展趋势。 根据招聘单位的地域、行业、职位等特点,找出哪些是热门地域、行业、职位等,并针对热门行业,分析其相应的技术要求;分析当前人才市场的需求现状并预测其未来短期的需求量。	1.文本聚类 2.时间序列预测	TipDM-TB R 语言 Python
基于网络新闻数据的食品安全挖掘	关于食品安全的网络新闻蕴含着大量的食品安全信息,根据历史食品新闻报道发现最受关注的食品供应环节,深度挖掘食品安全问题细节;根据历史食品新闻报道发现最受关注的食品种类,深入挖掘新闻对该类食品报道的主要关注点,进而获得该类食品安全问题的主要原因。	1.去空值、重复值、干扰词、 2.最大正向匹配法、分词、停用词过滤 3.关键词提取 4.关联词汇提取 5.K-Means 聚类、MIC、BT 6.语意网络图、雷达图、饼图	TipDM-TB R 语言 Python

基于垃圾短信的文本内容识别	垃圾短信已日益成为困扰运营商和手机用户的难题,严重影响到人们生活、侵害到运营商的社会形象以及危害着社会稳定。传统的基于策略、关键词等过滤的效果有限,很多垃圾短信“逃脱”过滤,继续到达手机终端。本案例 基于短信文本内容,结合机器学习算法、大数据分析挖掘来智能地识别垃圾短信及其变化。建立模型准确地、完整地识别出垃圾短信、正常短信。	1.欠抽样与过抽样 结合的类不平衡处理 2.支持向量机分类 3.精准率 4.召回率 5.F1 评价指标	TipDM-TB R 语言 Python
用户浏览新闻行为分析与智能推荐	挖掘用户的潜在兴趣并进行相应的新闻推荐,就能够产生更大的社会和经济价值。通过对带有时间标记的用户浏览行为和新闻文本内容进行分析,分析用户的新闻浏览模式和变化规律,基于新闻的文本内容,对新闻进行合理的分群;根据用户浏览新闻行为,实现为每个用户的新闻的个性化智能推荐。	1.正则匹配 2.文本分词 3.LDA 主题模型及可视化检验 4.物品协同过滤算法 5.时序加权推荐列表 6.精准率 7.召回率 8.F1 评价指标	TipDM-TB TipDM-HB R 语言 Python

消费者投诉举报信息意见挖掘	针对 12315 消费者投诉举报专线的投诉数据,从中挖掘出有用信息,一方面方便工商部门及时有效的受理和查处各类经济违法违章案件,维护市场经济秩序;另一方面可以更好的维护消费者权益;同时也可商家提供改进商品提高竞争力的依据。统计投诉、举报、咨询三类信息中的品牌分布,并绘制词云图;归纳消费者投诉的主要内容(方面),分析投诉原因。	1.词条-文档矩阵 2.高频词云图 3.LDA 主题分析 4.语义网络分析	TipDM-TB R 语言 Python
电信诈骗事件演化分析与挖掘	基于当前网络信息发展与电信诈骗事件的特点,从海量新闻文本中快速了解事件的发展过程及民众的情绪变化和关注焦点,找出事件的问题根源对于快速了解事件发展具有重要意义,也为相关部门更好地管控和处理事件提供参考依据。根据电信诈骗相关新闻文章,分析事件发展过程,深度挖掘事件问题根源;分析民众对该事件发展过程的情感变化,挖掘出民众关注的焦点问题。	1.结巴分词 2.TF-IDF 词频-逆文档频率 3.LDA 主题分析及基于相似度的主题数确定 4.情感倾向性分析 5.语义网络分析	TipDM-TB R 语言 Python

电子商务	法律网站 用户行为 分析与服 务推荐	根据某家法律网站采集到的用户访问数据，按地域研究用户访问时间、访问内容、访问次数等分析主题，深入了解用户对访问网站的行为和目的及关心的内容。借助大量的用户的访问记录，发现用户的访问行为习惯，对不同需求的用户进行相关的服务页面的推荐。	1.用 R 连接数据库取数据 2.数据库查询语句 3.正则匹配 4.协同过滤算法 5.相关系数：余弦相关系数、杰卡德相关系数	TipDM-TB TipDM-HB R 语言 Python
	基于关联 规则的菜 品智能推 荐	目前餐饮企业的点餐服务仅依赖于服务员的个人经验，但是对于多项菜品、以及它们的毛利率和店家的主推意愿，个人的经验就比较有限了，根据用户的历史点餐行为数据，挖掘菜品之间的关联度；借助菜品的关联度，结合业务理解，得出菜品推荐综合评分，完成菜品的智能推荐。	1.Apriori 关联规则 2.综合评分方法	TipDM-TB TipDM-HB R 语言 Python
	基于类不 平衡问题 的客户购 买品牌预 测	用户作为电商企业的主要经营对象，分析用户的行为记录，探索用户的行为规律，并将这些规律和网站经营策略相结合，对网站的营销方案作出有利的修改，以及为电商企业增加利润和实施高效管理有着重要意义。根据	1. 类不平衡问题处理 2. 随机森林建模 3. 混淆矩阵 4.精度 p 5.召回率 r	TipDM-TB R 语言 Python

		某电子商务网站部分用户的 4 个月的行为日志, 建立预测购买模型, 预测下个月用户会购买的品牌, 得出购买行为的一般特征。	6.F1 值	
	基于 LFM 的商品评分预测	商品评分是电子商务网站的一项必备环节, 若用户对商品的评分高, 则可以认为用户“喜欢”这件商品; 反之, 若用户对商品的评分较低, 则可认为用户“不喜欢”这件商品。根据某电商网站提供的数据进行商品评分预测。	1. 评分矩阵 2. LFM 算法 3. 评价指标 MAE、RMSE	TipDM-TB R 语言 Python
	竞赛网站用户行为分析与智能推荐服务	根据泰迪杯竞赛网站大量用户的访问数据, 发现用户的访问行为习惯, 对不同需求的用户进行相关服务页面的推荐; 根据用户访问内容、访问次数等属性特征, 深入了解用户对访问网站的行为和目的及关心的内容。	1. 分布分析 2. 缺失值插补 3. 数据变换 4. 数据去重 5. 杰卡德相关系数 6. 协同过滤算法 7. 推荐效果评价分析, 精确率、召回率。	TipDM-TB R 语言 Python

网站用户 浏览行为 分析及分 群研究	泰迪杯竞赛网致力于为用户提供丰富的泰迪杯竞赛信息、数据挖掘培训咨询服务,并为参赛者提供了往届优秀作品作为参考,为高校提供了丰富的教学资源。依据用户的历史浏览记录,分析用户的行为特征和兴趣偏好;根据用户的行为特征和兴趣偏好将用户划分成不同的群体,并分析各个群体的属性特征。	1. 分布分析 2. 缺失值插补 3.数据变换 4.数据去重 5. Kmeans 聚类 6. 用户特征图	TipDM-TB TipDM-HB R 语言 Python
O2O 优惠 券个性化 投放	对商家而言,滥发的优惠券可能降低品牌声誉,同时难以估算营销成本。因此,个性化投放是提高优惠券核销率的重要技术,它可以让具有一定偏好的消费者得到真正的实惠,同时赋予商家更强的营销能力。根据用户线下线上使用消费券的历史记录,预测用户在拿到一张优惠券后使用的概率。	1. 决策树 2. 混淆矩阵 3. 精度 p 4.召回率 r 5.F1 值	TipDM-TB R 语言 Python

金融保险	P2P 网络 信贷获贷 结果预测	金融机构建立信用卡风险管理体系，面临的重要问题是如何准确识别出哪些客户为高风险类客户，哪些客户为禁入类客户。基于某互联网企业提供的贷款申请数据，包括个人申请和企业申请。将申请数据分为个人申请客户和企业申请客户；分别建立个人申请和企业主申请获贷模型。	1.缺失值插补 2.字符型数据处理 3.SMOTE 数据不平衡处理 4.WOE 编码 5.逻辑回归	TipDM-TB R 语言 Python
	金融服务 机构资金 流入预测	某金融服务机构拥有上亿会员，并且业务场景中每天都涉及大量的资金流入和流出，面对如此庞大的用户群，资金管理压力会非常大。对货币基金而言，资金流入意味着申购行为，资金流出为赎回行为。根据用户的历史申购和赎回信息，预测蚂蚁金服次月每天的申购总额和赎回总额。	1.平稳性检验 2.白噪声检验 3.时间序列 BIC 图定阶 4.ARIMA 模型	TipDM-TB R 语言 Python
	信用卡高 风险客户 识别	为了推进信用卡业务良性发展，减少坏账风险，台湾各大银行都进行了信用卡客户风险识别相关工作，建立了相应的客户风险识别模型。判断识别出哪些客户为高风险类客户，哪些客	1.分布分析 2.关联分析 3.数据筛选 4.数据归一化 5.评分卡模型	TipDM-TB R 语言 Python

	户为禁入类客户。对不同客户类别进行特征分析，比较不同客户的风险。 评估该机构的信用卡业务风险，针对目前的情况提出风控建议。	6.K-Means 聚类	
P2P 网络 信贷风险 评估	风险控制主要目的是采取各种措施和方法，消灭或降低信用风险，流动性风险，市场风险等风险事件发生的各种可能性，或减少发生时造成的损失。做好风险控制，能够最大程度提高信贷资产收益，保证投资人的资金安全。找出影响用户逾期还款的关键因素；预测用户逾期还款的概率。	1.分布分析 2.多表合并 3.多重共线性分析 4.梯度提升方法 Gradient Boosting 5.数据降维	TipDM-TB R 语言 Python
贷款产品 申请审批 结果预测	根据网络贷款平台平台的历史数据，能预测金融机构回馈的审核结果。用户在提交贷款产品订单时能大概知道申请成功的概率，在一定程度上可以提高平台用户的使用体验。另一方面，平台方还希望知道影响贷款成功的关键因素，以便给申请贷款的用户更合适的建议。	1.分类变量的处理 2.缺失值处理 3.随机森林算法 4.GBM 算法	TipDM-TB R 语言 Python

交通运输	航空公司客户价值分析	航空市场竞争激烈,某航空公司面临着常旅客流失、竞争力下降、航空资源未充分利用等经营危机。通过积累的大量的会员档案信息和其乘坐航班记录,建立合理的客户价值评估模型对客户进行分群,分析比较不同客户群的客户价值,对不同价值的客户类别提供个性化服务,并制定相应的营销策略。	1.零-均值标准化 2.一致性分析 3.回归插值法 4.K-Means 聚类分析	TipDM-TB TipDM-HB R 语言 Python MATLAB
	城市公交站点设置的优化分析	本案例主要以城市公交车的刷卡数据和 GPS 车载数据为原始数据。采用 DESCAN 密度聚类法对 GPS 数据进行聚类分析区分公交站点;运用概率论模型计算得出下车人数。最终得出 OD 矩阵,推出居民出行规律,进而对公交站点设置提出优化建议。	1. 均值插值法 2. DESCAN 密度聚类法 3. 泊松分布 4. OD 矩阵	TipDM-TB TipDM-HB R 语言 Python
	基于 SARIMA 混合模型的铁路站点客流量预测	铁路部门为了保持市场的竞争力,实现利润的最大化,需要了解日常铁路客运流量、淡旺季变动指数、冷热门线路的具体情况。按车次、时段(小时)、车站、区间(两个车站之间)等分析客流规律。考虑相关因素的影	1.分布统计 2.数据变换 3.函数构造 4.ARIMA 模型 5.SARIMA 模型	TipDM-TB R 语言 Python

	响, 构建客流量预测模型, 并预测未来两周的客流量。针对特定的区段, 优化设计车辆配置以及提出车站停靠方案。		
交通事故成因分析	道路交通具有动态性, 随机性, 因果性和再现性等特点, 这就直接导致了提高道路交通安全的复杂性。利用交通事故相关的多维度数据, 通过对事故类型、事故人员、事故车辆、事故天气、驾照信息、驾驶人员犯罪记录数据以及其他和交通事故有关的数据进行深度挖掘, 形成交通事故成因分析的模型。	1.关联度 2.格兰因果 3.灰色预测 4.卡尔曼滤波	TipDM-TB R 语言 Python
航空机场无线网络的设备连接数预测	航空机场无线网络的连入数量, 位置等情况在某个程度上反映了机场客流量的多少, 以及分布情况。根据某机场提供海量机场 WIFI 数据及安检登机值机数据, 预测未来指定的时间窗口内, 航空机场内每个无线网络点每单位时间内的平均设备连接数量。	1.分布统计 2.相关度分析 3.ARIMA 模型	TipDM-TB R 语言 Python

信息安全	网络入侵 自动识别	入侵检测是网络安全领域中一个较新的课题,检测引擎作为入侵检测系统的核心模块,其检测速度快慢直接影响网络入侵检测系统的效率,模式匹配是入侵检测系统的重要检测方法,其性能对入侵检测系统至关重要。利用网络协议的高度规则性,采用协议分析的方法,结合模式匹配算法,归纳出入侵的关键特征,构建入侵的识别模型。	1.协议分析的方法 2.模式匹配算法 3.BP 神经网络	TipDM-TB R 语言 Python
	监控场景 下的行人 精细化识别	海量的视频监控流使得发生突发事件后,需要耗费大量的人力物力去搜索有效信息。行人作为视频监控中的重要目标之一,根据监控场景下多张带有标注信息的行人图像,完成图像数据中行人属性(头部、上身、下身、脚、帽子、包)的位置定位;研究行人精细化识别算法,自动识别出行人图像中行人的属性特征。	1.灰度化、二值化 2.canny 获取边缘 3.特征提取 4.图像形态学变换 5.图层识别	TipDM-TB R 语言 Python

	广告监测中的流量作弊识别	与传统的电视广告、户外广告采买相比,流量作弊一直以来被看作互联网广告特有的弊病。随着网络数据技术的发展进步,流量作弊也呈现出规模化、机器化、产业化的趋势。基于 IP, cookie, 设备 ID, 访问时间序列, UA 信息分布等行为属性来建立一个模型,区分正常用户曝光记录与作弊行为记录,并进行标记。	1、分类数据概率转换 2、缺失值处理 3、神经网络算法 4、混淆矩阵评价模型	TipDM-TB R 语言 Python
政务民生	基于数据挖掘技术的市财政收入分析预测	如何有效的利用地方财政收入,合理的分配,来促进地方的发展,提高市民的收入和生活质量是每个地方政府需要考虑的首要问题。根据广州市 1994-2013 年财政收入以及相关因素的数据,梳理影响地方财政收入的关键特征,分析、识别影响地方财政收入的关键特征的选择模型;对广州市 2015 年的财政总收入及各个类别收入进行预测。	1.Lasso 法 2.Adaptive-Lasso 变量选择法 3.灰色预测与神经网络组合模型	TipDM-TB R 语言 Python

	民政低保申请用户识别	低保制度还不完善,具体的实施也就存在诸多漏洞。针对不同种类低保户,提供他们最需要的救助是亟待解决的另一个难题。找出低保户与非低保户的特征,通过动态识别的方式,自动识别符合低保条件或者不符合低保条件的对象。	1.神经网络 2.SMOTE 算法 3.过抽样 4.ROC 曲线 5.精确率 6.召回率	TipDM-TB R 语言 Python
	基于数据挖掘技术的民政低保用户精准扶贫	由于缺乏科学的瞄准机制导致贫困原因不明,扶贫情况不明等,严重阻碍了扶贫工作的开展。利用数据挖掘知识精准的找出每个地区致使其贫困的主要原因。针对找出的原因对不同地区采取差异化的扶贫政策,使扶贫效果更加有效。	1.决策树 2.神经网络 3.k-means 聚类 4.雷达图	TipDM-TB R 语言 Python
其他行业	基于水色图像的水质评价	由于水色能反映水中浮游植物的种类和多少,可通过观察水色变化来调控水质,维持养殖水体生态系统合理的动态平衡。通过历史水产专家经验判断结果以及用数码相机按照标准进行水色采集的数据,利用图像处理技术,实现基于水色图像的水质自动评价。	1.图像切割 2.颜色矩阵 2.SVM 算法 3.混淆矩阵	TipDM-TB TipDM-HB R 语言 Python

多种农产品价格的分类智能预测	农产品价格是否合理,不仅影响农业生产的发展,农产品的流通、消费和农民的收入水平,而且影响工业品的成本和价格,影响国家同农民之间、城乡人民之间以及农民内部的物质利益关系,对整个社会经济生活的安定也关系重大。通过大量历史数据的分析,预测农产品未来的价格走势。	1.映射处理 2.样条插值法 3.ARIMA 模型 4.SARIMA 模型	TipDM-TB R 语言 Python
家用电器用户行为分析及事件识别	家用电器在使用过程中,因地区气候、区域不同、用户年龄性别差异,形成不同的机组使用行为,称之为用户的使用习惯。按地域研究用户访问时间、访问内容、访问次数等分析主题,深入了解用户对访问网站的行为和目的及关心的内容。借助大量的用户的访问记录,发现用户的访问行为习惯,对不同需求的用户进行相关的服务页面的推荐。	1.数值变化 2.特征提取 3.BP 神经网络	TipDM-TB TipDM-HB R 语言 Python

	应用系统 负载分析 与磁盘容 量预测	与应用系统关联的任何一种资源负载过大,都可能会引起系统性能下降甚至瘫痪。通过分析磁盘容量相关数据,预测应用系统服务器磁盘空间是否满足系统健康运行的要求。根据用户需求设置不同的预警等级,将预测值与容量值进行比较,对其结果进行预警判断,为系统管理员提供定制化的预警提示。	1.数据去重 2.ARIMA 模型 3.AIC 准则和 BIC 信息准则 4.平均绝对误差 5.均方根误差 6.平均绝对百分误差 7.使用容量定级	TipDM-TB R 语言 Python
	管网漏损 及居民用 户用水异 常分析	居民用水管道爆裂、渗漏,私接公共管道,水表故障(或人为调整)引发水量反常等异常用水现象频发,尤其是居民室内漏水以及盗取民用水为商业用水行为。归纳不同用户的用水行为特征,构建模型对用户进行分类,识别异常用水用户群。进一步对管网漏损进行检测定位,确定漏损发生的时间。	1. 箱线图 2.统计分布图 3. 均值插值法 4. C4.5 决策树算法 5. 混淆矩阵 6.ROC 曲线	TipDM-TB R 语言 Python

说明: 案例库不断更新中……

2 案例构成

数据挖掘教学实训案例主要由以下几部分构成：

案例介绍：

- 案例背景
- 提出挖掘目标
- 分析方法与过程
- 上机实践
- 拓展思考

配套资源：

- ☐ 源数据及过程数据
- ☐ 建模环境下的程序/模型
- ☐ 教学配套PPT
- ☐ 微课视频

3 案例示例

防窃漏电智能诊断系统通过采集电量异常、负荷异常、终端报警、主站报警、线损异常等信息，通过“失压失流”、“超合同容量用电”等用电报警信息，以及根据报警事件发生前后客户计量点有关的电流、电压、负荷数据情况来构建用户异常分析模型，从而实现检查客户是否存在窃电、违章用电等目的。本项目的成果已广泛应用于广东电网各地市单位，大幅度提高了防窃漏电诊断的效率及准确性，有效发现并查处窃漏电事件70余起，追回电量损失（含违约相关费用）数百万元，已累计产生经济效益超数千万元。

一种自适应的防窃漏电诊断方法，曾获广东电网公司科学技术发明专利一等奖（D2015ZL105-G8）和中国南方电网公司科学技术发明专利一等奖（2015-1-03-G08）。

3.1 案例介绍

3.1.1 背景与挖掘目标

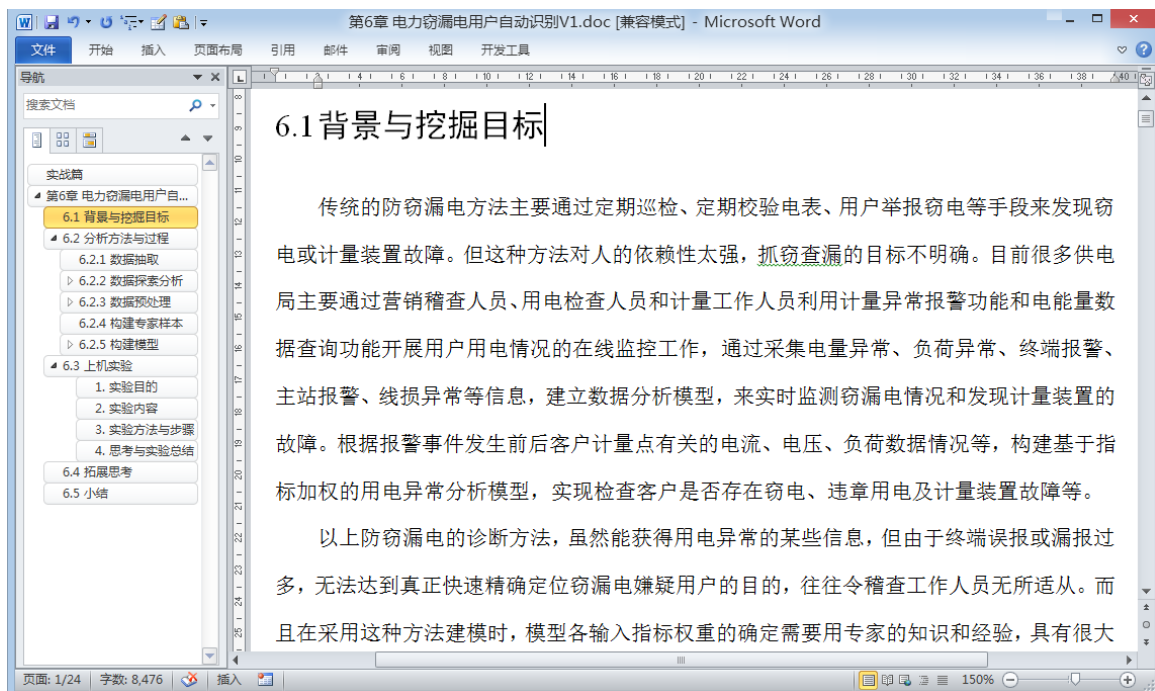


图 3-1 案例背景与挖掘目标文档

3.1.2 分析方法与过程

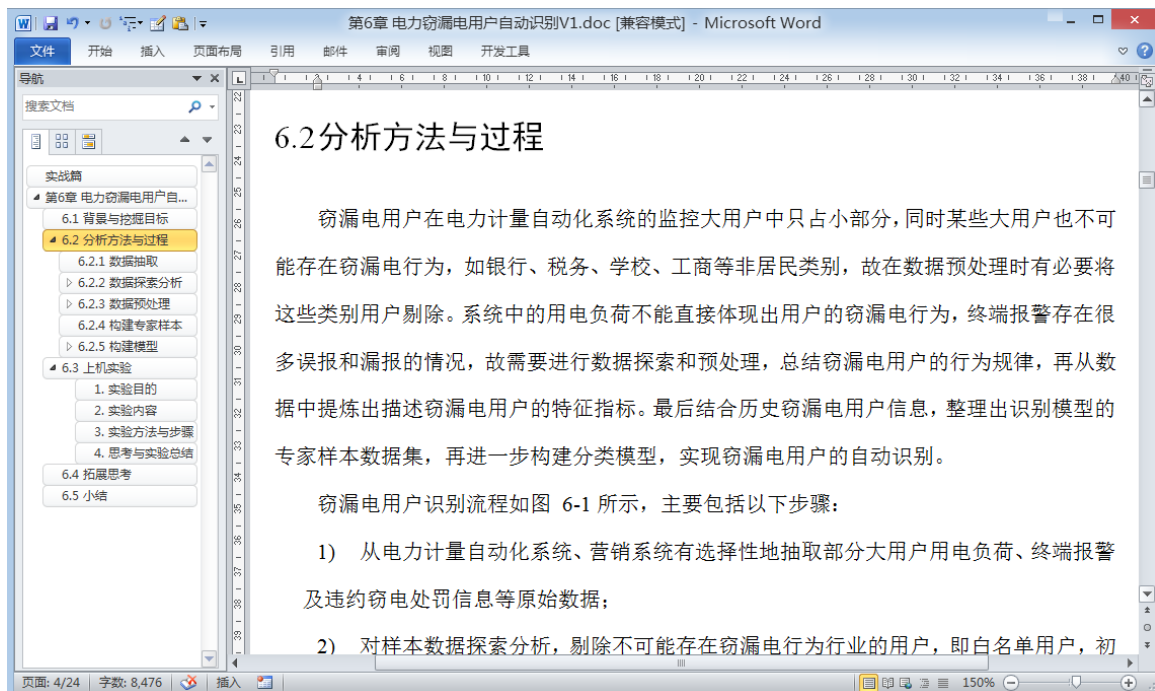


图 3-2 案例分析方法与过程文档

3.1.3 上机实践

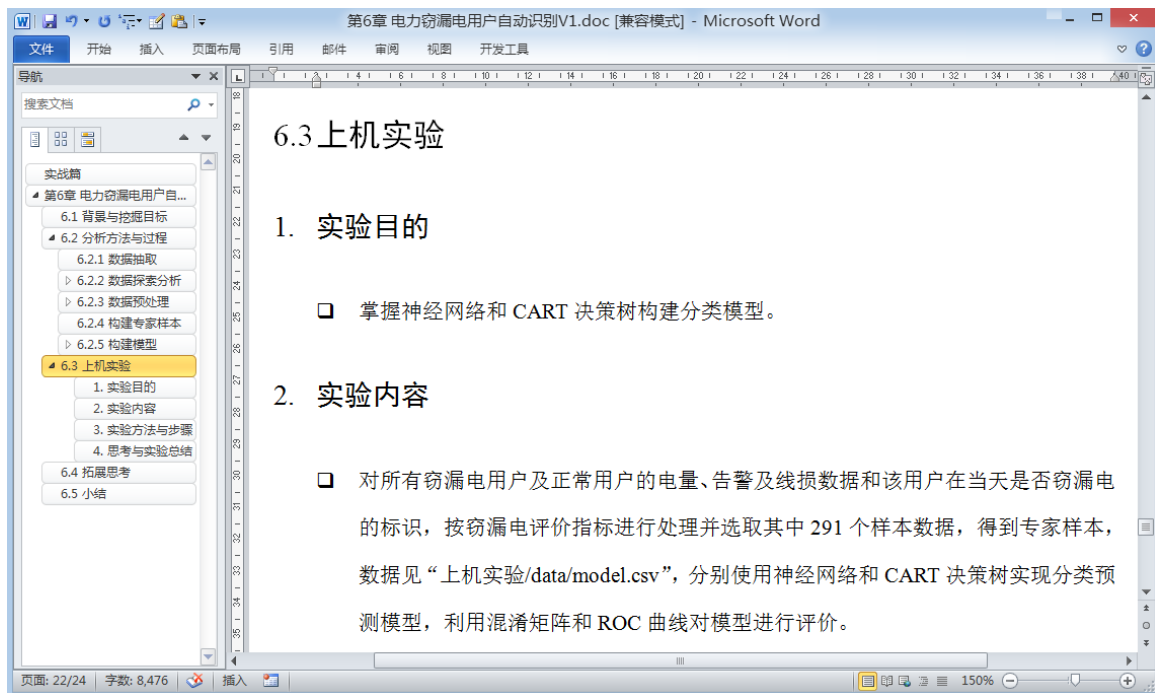


图 3-3 案例上机实验文档

3.1.4 拓展思考

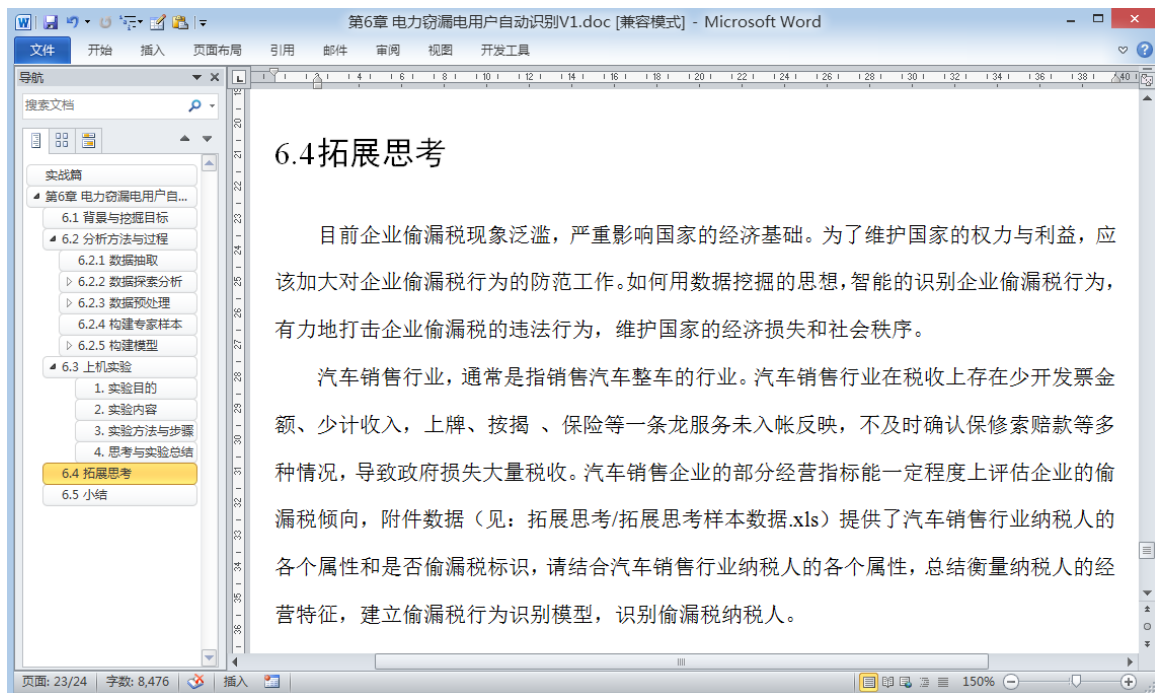


图 3-4 案例拓展思考文档

3.2 配套资源

案例配套资源包括：源数据及过程数据，常用建模环境下的程序/模型、教学配套PPT 及案件微课视频。

3.2.1 配套程序

名称	修改日期	类型	大小
bin2ind.m	2015/3/6 18:29	M 文件	1 KB
cal_successrate.m	2015/3/6 18:29	M 文件	1 KB
decision_tree.m	2015/3/6 18:29	M 文件	2 KB
decision_tree_test.m	2015/3/6 18:29	M 文件	1 KB
ind2bin.m	2015/3/6 18:29	M 文件	1 KB
lm_net.m	2015/3/6 18:29	M 文件	2 KB
plotroc_confusion.m	2015/3/6 18:29	M 文件	1 KB
ployinterp_column.m	2015/3/6 18:29	M 文件	2 KB
polyinterp.m	2015/3/6 18:29	M 文件	1 KB
preprocess_interp.m	2015/3/6 18:29	M 文件	1 KB
saveconfusion_roc.m	2015/3/6 18:29	M 文件	1 KB
split2train_test.m	2015/3/6 18:29	M 文件	1 KB
test_decisontree_lm.m	2015/3/6 18:29	M 文件	1 KB
test_net.m	2015/3/6 18:29	M 文件	1 KB
test_preprocess_interp.m	2015/3/6 18:29	M 文件	1 KB
train_net.m	2015/3/6 18:29	M 文件	1 KB

图 3-5 案例配套程序 (MATLAB/R/Python)

3.2.2 配套数据

名称	修改日期	类型	大小
missing_data.xls	2015/3/6 18:29	Microsoft Office...	25 KB
model.xls	2015/3/6 18:29	Microsoft Office...	40 KB
test_model.xls	2015/3/6 18:29	Microsoft Office...	26 KB
train_model.xls	2015/3/6 18:29	Microsoft Office...	36 KB

图 3-6 案例配套原始数据

3.2.3 教学配套 PPT



图 3-7 教学配套 ppt



图 3-8 教学配套 PPT

3.2.4 微课视频

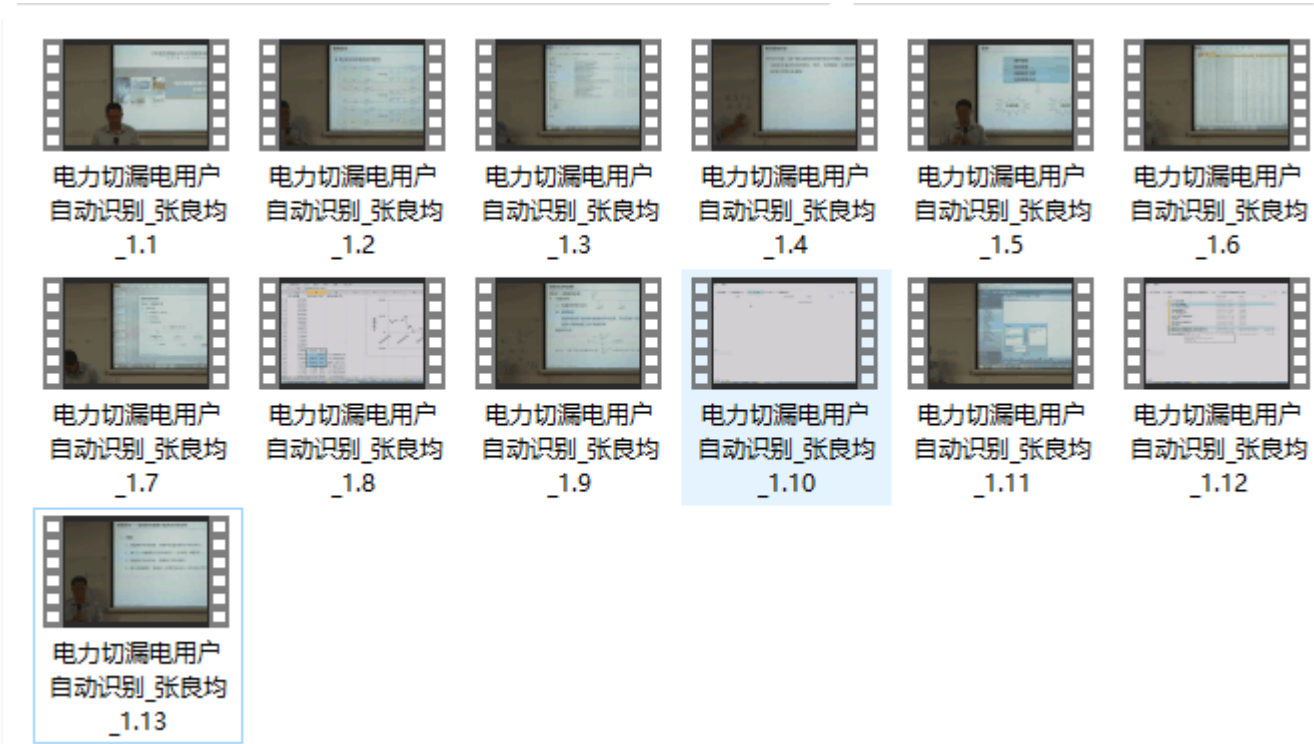


图 3-9 案例配套微课视频