第九章 员工舞弊识别
-机器学习

目录

CONTENT

章前导读

舞弊是企业内部治理的顽疾，正日益成为全球性的焦点问题。据美国注册舞弊审查师协会（ACFE）发布的《2022年ACFE全球舞弊调查报告》（以下简称《报告》），舞弊给包括政府和企业在内的各类组织带来的经济损失约为其全年总收入的5%，这一数据与20年数据持平，每起案件的平均损失为 1,783,000 美元，这一数据与20年相比有18%的上升，说明舞弊损失的影响有一定扩大。舞弊不仅使企业蒙受直接损失与处罚损失，还导致严重的声誉损失，对企业的资本市场与产品市场均造成不利影响，严重损害企业价值、阻碍企业发展。

01

反舞弊控制手段

02

《报告》指出81%的受害组织在发现舞弊之后调整了反舞弊手段，其中：

75%的组织增加管理层审核的流程；

64%的组织增加主动的数据监测与分析；

54%的组织增加突击审计；

48%的组织增加内审部门；

42%的组织增加反舞弊培训。

03

控制效果

据统计，最高效的舞弊行为控制手段是“主动的数据监测与分析”，可缩短56%的时间快速发现舞弊行为；其次是突击审计和轮岗，可缩短50%的时间。轮岗可有效减少54%的舞弊损失，举报热线和突击审计可有效减少50%损失，主动的数据监测与分析可有效减少47%的损失。

第一节

机器学习技术基础

机器学习的概念

机器学习是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。机器学习专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。机器学习是人工智能的核心，是使计算机具有智能的根本途径。它的应用已遍及人工智能的各个分支，如专家系统、自动推理、自然语言理解、模式识别、计算机视觉、智能机器人。

学习能力是智能行为的一个非常重要的特征

学习能力是智能行为的一个非常重要的特征，机器学习在人工智能的研究中具有十分重要的地位。一个不具有学习能力的智能系统难以称得上是一个真正的智能系统，但是以往的智能系统都普遍缺少学习的能力。例如，它们遇到错误时不能自我校正；不会通过经验改善自身的性能；不会自动获取和发现所需要的知识；它们的推理仅限于演绎而缺少归纳，因此至多只能够证明已存在的事实、定理，而不能发现新的定理、定律和规则等。随着人工智能的深入发展，这些局限性表现得愈加突出。正是在这种情形下，机器学习逐渐成为人工智能研究的核心之一。

机器学习的研究是根据生理学、认知科学等对人类学习机理的了解，建立人类学习过程的计算模型或认识模型，发展各种学习理论和学习方法，研究通用的学习算法并进行理论上的分析，建立面向任务的具有特定应用的学习系统。这些研究目标相互影响、相互促进。

机器学习的分类

从学习策略看，机器学习可分为机械学习、示教学习、演绎学习、类比学习、基于解释的学习、归纳学习等。从所获取知识的表示形式看，机器学习可分为符号表示类和亚符号表示类。决策树、形式文法、产生式规则、形式逻辑表达式、框架和模式等属于符号表示类；而代数表达式参数、图和网络、神经网络等则属亚符号表示类。 从学习形式看，机器学习可分为监督学习、无监督学习、半监督学习和强化学习。下面，从这个角度来简要介绍一下。

监督学习

监督学习是从给定的训练数据集中学习出一个函数，当新的数据到来时，可以根据这个函数预测结果。监督学习的训练集要求包括输入和输出，也可以说是特征和目标。训练集中的目标是由人标注的。比如，我们想让计算机自动识别出垃圾邮件，就需要我们收集大量的垃圾邮件和非垃圾邮件数据，并告诉计算机哪些是垃圾邮件，哪些不是垃圾邮件，即给我们收集的数据打标签后再输入计算机，让计算机经过训练后可以识别新输入的样本是否为垃圾邮件。监督学习主要包括回归分析和统计分类两类算法，比较经典的算法有线性回归、逻辑回归、支持向量机、决策树、朴素贝叶斯等。

机器学习的分类

无监督学习

无监督学习与监督学习相比，训练集没有人为标注的结果，主要用于从未标注数据集中挖掘相互之间的隐含关系。那么，计算机如何对没有标签的数据进行分类呢？举一个例子，在一个二维空间中有一些样本点，我们不知道这些样本点的数据类别，这就需要我们假设一个条件：在空间中相聚更近的点即为一类，如果这个假设成立，我们就可以根据样本的空间信息，设计算法将其聚集为两类或多类，从而实现没有标签的机器学习分类。常见的无监督学习算法有聚类、主成分分析、潜在语义分析等。

半监督学习

半监督学习介于监督学习与无监督学习之间，即输入的数据一部分有标签，一部分无标签。半监督学习的运用非常广泛，随着互联网的不断发展，数据量不断增大，进行数据标签的成本也越来越大。因此，利用少量标注数据和大量没有标注的数据训练一个更好的机器学习算法，成为了机器学习领域的热点之一。大多数半监督学习算法是无监督式和监督式算法的结合。

强化学习

强化学习通过观察来学习做出什么动作，每个动作都会对环境有所影响，学习对象根据观察到的周围环境的反馈来做出判断。其原理为，计算机通过与外界环境交互获得数据，随机的产生行为并获得该行为的结果，程序需要通过定义这些行为的收益函数，对行为进行奖励或者惩罚，同时，我们需要设计算法让计算机自动改变自己的行为模式来最大化收益函数，使其行为达到一个最佳效果。例如，计算机下棋，如果下赢了，我们就进行奖励；如果下输了，我们就进行惩罚。经过多次训练，计算机就可以成为一个下棋高手。强化学习在游戏、自动驾驶、推荐系统等领域有着广阔的应用前景。

大数据审计分析的机器学习实现

一、审计应用分析的机器学习实现

- 在风险评估阶段，审计人员可以通过使用机器学习工具，充分利用内外部数据，更好地理解企业数据，帮助审计人员侦测到目前可能被忽视的问题，快速发现被审计单位与行业周期不一致的地方，以便在后续审计程序中予以重点关注，大大降低审计风险。

- 在实质性程序阶段，对重点关注领域，可以由抽样审计转变为全量审计，应用机器学习模型对大量同类交易数据进行异常值检测，从而锁定审计疑点。审计人员还可以通过比较实际数据和机器学习生成的预测数据来检测异常。审计人员可以将机器学习模型提供的会计估计预测值作为一个判断基准，与管理层提供的估计值进行比较，以便确定需要进一步调查的领域。

- 审计人员可以利用关联规则等数据挖掘与分析方法，对被审计单位的各项业务进行数据分析，发现审计线索，提示审计风险。例如，可以对企业银行账户的收支明细数据进行分析，可能发现某一收方账户名称和某一付方账户名称总是同时出现，则可对被审计单位与该收方账户的全部业务及该付方账户的全部业务进一步调查分析，判断是否存在虚假交易的风险。

- 在内部审计中，审计人员可以依据历史经验，整理全部员工费用报销单据信息，梳理违规报销单据的N个特征，应用朴素贝叶斯算法，训练判断费用报销单据是否存在违规的模型，并将其内嵌到企业的报销系统中，用来自动识别出可能存在违规报销的单据，并推送消息给审计人员，以便审计人员做进一步核查。

大数据审计分析的机器学习实现

二、机器学习审计应用实例

➢ **智能文档审阅平台**

德勤（Deloitte）的Argus是基于机器学习的智能文档审阅平台。Argus利用计算机认知技术，将原本需要耗费大量人工执行的文档审阅工作，交由平台进行智能化大批量处理。Argus基于先进的自然语言处理和机器学习技术，不断通过与操作人员的互动持续学习，将原本耗时、手动的作业，更快速、轻松地完成。作为德勤全球在英文文档领域的智能平台，Argus在发布之后的三年时间内就帮助了德勤上千位专业人员审阅了几十万份的诸如各类合同、票据、会议纪要、法律文书的文档，并快速准确地提供分析结果，大大提高了德勤专业人员的工作效率。德勤中国针对国内业务需求，在整合Argus英文文档审阅功能的基础上，成功研发出一个综合性、多语种支持的智能文档审阅平台“IDRP”，引领国内的专业文档处理进入人工智能时代。

➢ **数据审计、分析工具**

普华永道（PwC）基于机器学习技术开发的Halo数据审计工具，能够测试海量关键业务数据。这有助于审计员分析样本总体，完善风险评估、分析和测试，并提升审计人员的洞察力，加深审计人员对客户业务的了解，从而提升审计质量。Halo充分发挥数据力量，不断革新普华永道的审计方式。结合趋势分析和比率分析，审计人员可以在审计中利用已获得的信息向客户提供深刻的见解。各类图表使客户的潜在高风险交易一目了然，可加深审计人员对客户业务的理解、开展更加有效的对话。过去需人工完成的任务现已实现自动化，效率大幅提升。

大数据审计分析的机器学习实现

二、机器学习审计应用实例

- **数据审计、分析工具**

安永（EY）持续研发和应用的数据分析工具EY Helix，基于机器学算法设计了大量的数据分析器，帮助审计人员加深了对被审计单位业务的理解，使他们更全面地了解企业的方方面面，识别新交易流，评估会计信息质量，聚焦值得关注的风险，识别流程中的趋势和异常情况，及时制定有针对性的审计策略。通过使用EY Helix，可以真正实现数据分析驱动审计，审计师可以更容易地确定关键审计事项，并将时间和精力集中在关键事项的审计查证上。

- **智能组合工具**

毕马威（KPMG）的智能组合工具Ignite，整合了自然语言处理、机器学习、深度学习和光学字符识别等功能，旨在将语音和图像等非结构化数据转换为有意义的编码结构后进行快速分析，以应对复杂的业务挑战。该工具可提供30多个人工智能服务和组件，包括商业贷款文档信息的自动识别、提取和评估，证券招股说明书的自动审查和评估，供应商及客户合同的自动审查、分析等。毕马威已应用该工具提供了超过100项服务，处理了超过7亿页的文档。

第二节

实战演练-员工舞弊识别

实训任务

新道物流集团总部内部审计部，正在对集团各子公司进行2021年度财务审计。在进行总体分析时，审计部发现，河北分公司物流成本中的油料费占总成本比、及占收入比，均明显高于集团平均水平，故决定对河北分公司该项成本费用进行进一步调查分析。作为该项目的负责人，部门经理指派你制定调查分析计划，查明原因，识别确认是否存在员工舞弊情况。

知识背景-舞弊的内涵

注册舞弊审查师协会（Association Of Certified Fraud Examiners，ACFE）成立于1988年，由Joseph T.Wells博士（CFE、CPA）创立，是世界上最大的反舞弊组织，也是反舞弊培训和教育的内容和考试提供方。ACFE与150多个国家的85000多名会员一起，正在全球范围内努力减少商业舞弊，并持续提供更有效打击舞弊所需的培训和各种资源。

Q1 什么是舞弊?

根据注册舞弊审查师协会（ACFE）的定义，广义的会计舞弊行为分为三大类别：腐败、挪用资产和财务报表舞弊。美国《国家审计准则第82号通知》对舞弊的定义是：为了得到他人的信任，故意歪曲事实真相，并且明知其行为是违法的或者错误的，舞弊者因此行为获得利益，同时第三者因此行为造成损失。我国审计准则《财务报表审计中对舞弊的考虑》对舞弊的界定：舞弊是指被审计单位的管理层、治理层、员工或第三方使用欺骗手段获取不正当或非法利益的故意行为。

Q2 什么是员工舞弊?

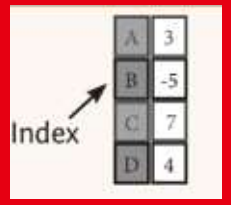
员工舞弊，是指公司内部的雇员，以欺骗性的手段不正当地获取组织的钱财或其他财产的行为。如贪污，受贿，以权谋私，挪用资金等行为。员工舞弊又称非管理舞弊，其舞弊者为公司员工，而公司为受害者。区别于管理舞弊，管理舞弊是指管理当局故意通过具有严重误导性质的财务报表损害投资者和债权人的行为，也称为财务报告舞弊。

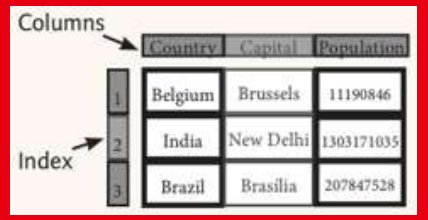
知识背景- Pandas库

作用

Pandas（https://pandas.pydata.org/）提供了高级数据结构和函数，这些数据结构和函数的设计使得利用结构化、表格化数据的工作快速、简单、有表现力。

Series





DataFrame

Pandas将表格和关系型数据库（例如SQL）的灵活数据操作能力与NumPy的高性能数组计算的理念相结合。它提供复杂的索引函数，**使得数组的重组、切块、切片、聚合、子集选择更为简单**。由于数据操作、预处理、清洗在数据分析中非常重要，因此本次实训应熟练掌握pandas库的这些应用技能。

优势

知识背景-PyOD库

PyOD库简介

PyOD库是一个异常点检测算法工具库。异常检测（又称outlier detection、anomaly detection，离群值检测）是一种重要的数据挖掘方法，可以找到与“主要数据分布”不同的异常值（deviant from the general data distribution），比如从信用卡交易中找出诈骗案例，从正常的网络数据流中找出入侵，有非常广泛的商业应用价值。同时它可以被用于机器学习任务中的预处理（preprocessing），防止因为少量异常点存在而导致的训练或预测失败。

优势

- 包括近20种常见的异常检测算法，以及最新的深度学习，如对抗生成模型和集成异常检测；
- 支持不同版本的Python（2.7和3.5+），支持多种操作系统（windows，macOS和Linux）；
- 简单易用，只需要几行代码就可以完成异常检测；
- 使用JIT和并行化进行优化，加速算法运行及扩展性，可以处理大量数据。

知识背景-KNN算

KNN算法简介

KNN （K-NearestNeighbor）算法，即邻近算法，或者说K最邻近分类算法，是数据挖掘分类技术中最简单的方法之一。所谓K最近邻，就是K个最近的邻居的意思，说的是每个样本都可以用它最接近的K个邻近值来代表。近邻算法就是将数据集合中每一个记录进行分类的方法。

算法流程

①准备数据，对数据进行预处理 ；

②计算测试样本点（也就是待分类点）到其他每个样本点的距离；

③对每个距离进行排序，然后选择出距离最小的K个点；

④对K个点所属的类别进行比较，根据少数服从多数的原则，将测试样本点归入在K个点中占比最高的那一类。

算法优缺点

优点：KNN方法思路简单，易于理解，易于实现，无需估计参数。

缺点：该算法在分类时有个主要的不足是，当样本不平衡时，如一个类的样本容量很大，而其他类样本容量很小时，有可能导致当输入一个新样本时，该样本的K个邻居中大容量类的样本占多数。该方法的另一个不足之处是计算量较大，因为对每一个待分类的文本都要计算它到全体已知样本的距离，才能求得它的K个最近邻点。

需求分析-业务需求分析

审计客体

新道物流河北分公司

审计对象

2021年“物流成本-油料费”

制定调查分析计划

要收集哪些资料？
要从哪几个角度分析？
输出成果是什么？

执行调查分析

在执行的过程中要应用哪些技术？

审计目标

查明2021年“物流成本-油料费”高的原因，识别是否存在员工舞弊

需求分析-业务需求分析

收集资料

- 政策信息：集团实行“一车一卡”政策，以子公司为单位开设一张主卡、每辆车一个副卡。每月由财务部向主卡“充值”，油卡保管员从主卡额度账户向副卡备付金账户分配金额，使用时由驾驶员持卡加油；
- 油料费的详细数据：《加油交易信息表》；
- 车辆的详细数据：《车辆信息表》。

分析角度

- 加油交易的频次是否异常？
- 日加油次数是否合理？
- 次加油量是否合理？
- 加注油品是否正常？

输出成果

- 日加油次数预警表；
- 超容积加油预警表；
- 加注其他油品预警表；
- 异常加油检测报告。

需求分析-技术需求分析

已收集资料

政策信息

集团实行“一车一卡”政策，以子公司为单位开设一张主卡、每辆车一个副卡。每月由财务部向主卡“充值”，油卡保管员从主卡额度账户向副卡备付金账户分配金额，使用时由驾驶员持卡加油。

车辆信息

车辆信息表

序号	车辆类别	车辆品牌	车牌号	卡号	装备日期	加油种类	车辆油箱容积(L)	保管人	备注
1	运输车辆	雷诺	冀C.10972	1000113700009582101	2017-07-27	汽油	70	贾萌	运输一队
2	运输车辆	东风	冀C.50551	1000113700010676225	2017-01-12	汽油	100	丁军	运输一队
3	运输车辆	斯泰尔王	冀C.50822	1000113700010676228	2019-03-26	汽油	70	刘飞	运输一队
4	运输车辆	豪沃	冀C.50176	1000113700010676229	2017-09-06	汽油	100	李兆林	运输一队
5	运输车辆	豪沃	冀C.50189	1000113700010676232	2018-11-26	汽油	60	刘涛	运输一队
6	运输车辆	优迪狮	冀C.5142D	1000113700010676236	2017-12-14	柴油	70	谭宗邦	运输一队

加油信息

中国石化加油IC卡台账对账单

客户名称: 新道物流股份有限公司河北分公司　　客户编码 370200170268

网点名称:　　账单类型 卡账

起止时间: 2021-01-01----2021-12-31　　应用类型 电子油票

操 作 员: 肖俊　　打印日期 2022-3-31

卡号	时间	业务类型	品种	数量	单价	金额	奖励分值	优惠价	余额	地点	操作员	备注
1000113700009582101	2021-1-18 17:08	加油	95号车用汽油VI	28.17	7.1	200.0	0.0	7.1	2495.95	石家庄42站	高庆朋	
1000113700009582101	2021-1-21 16:33	加油	95号车用汽油VI	63.36	7.63	483.44	0.0	7.63	2012.51	石家庄42站	高庆朋	
1000113700009582101	2021-1-24 16:27	加油	95号车用汽油VI	19.8	7.17	141.97	0.0	7.17	1870.54	石家庄06站	高庆朋	
1000113700009582101	2021-1-27 11:01	加油	95号车用汽油VI	35.35	7.68	271.49	0.0	7.68	1599.05	石家庄42站	高庆朋	
1000113700009582101	2021-1-30 10:40	加油	95号车用汽油VI	53.92	7.89	425.43	0.0	7.89	1173.62	石家庄06站	高庆朋	
1000113700009582101	2021-2-3 13:34	加油	95号车用汽油VI	56.72	7.6	431.07	0.0	7.6	742.55	石家庄42站	高庆朋	