

智慧审计的七种武器之复杂数据分析

(来源：德勤微信公众号，2019-06-04)

一、前言

数据分析已广泛应用于各个领域，很多智能的、深入的、有价值的信息只有通过深入的数据分析才能获得。毫无疑问，在内部审计中数据分析同样占据越来越重要的地位。随着企业集成的数据源不断增加、数据量不断累积、数据更新频率不断加快，传统的数据存储、处理以及分析技术已捉襟见肘。因此，复杂数据分析能力就显得尤为重要，它直接影响了最终能否获得有价值的信息。在大数据时代，内部审计要想有效的履行企业风险控制第三道防线的职能，需紧跟大数据技术的发展步伐，提升复杂数据分析能力。

二、正文

1、形态

复杂数据分析 (Sophisticated Data Analysis)，是通过整合各种功能强大的数据处理工具，以高性能的数据处理方式来采集、整合、分析复杂数据，并快速从中挖掘出有效信息的能力。企业数据通常来自于内外部的多个数据源，具有体量大、维度多、更新快、价值密度低和数据形式多样的特点。其面对的数据库有传统的关系型数据库如MySQL、Oracle 等，也有适用于处理大量数据的高访问负载以及日志系统的键值数据库、适用于分布式大数据管理的列存储数据、适用于Web应用的文档型数据库和适用于社交网络的图形数据库等非关系型数据库。可处理的数据类型分为结构化数据，半结构化数据和文本、

图像、声音、影视等非结构化数据。

大量的、复杂的数据增加了处理的难度，但同时提供了更大的信息量。复杂数据分析可以从大量数据中发现其蕴含的模式和规律，进而产生更多的价值。

2、招式

复杂数据分析所涉及到的技术包括数据采集、数据处理、统计分析和数据建模。



数据采集

复杂数据分析的数据源可以是传统的内部数据库也可以是来自网络上的外部数据。内部数据采集可使用 SQL 将内部关系型数据库中的数据提取出来，或使用 NoSQL 将分布的、异构数据源中的数据文件如图片、文本等抽取出来。外部数据采集通常会用到网络爬虫技术，从 web 中获取所需的海量数据，如政府、研究机构或新闻网站上发布的信息等。数据获取后，对其进行清洗、转换、集成，最后加载到数据仓库或数据集市中使用。

数据处理

由于复杂数据分析所面对的数据通常体量巨大且形式多样，一般需要使用到更高性能的计算架构和存储系统。例如在处理用户 app 浏

览记录等体量巨大的数据时使用分布式计算的 MapReduce、Spark 计算框架可以提升计算能力应对更复杂的数据并减少数据处理时间，使用分布式文件存储 HDFS 进行大规模数据协同工作来提升数据的吞吐能力和速度。针对非结构化的数据，往往需要对数据进行深入的理解，并通过复杂数据分析将结构多样、语义多样的非结构化数据进行结构化处理，提取出可以直接进行分析的数据。例如针对半结构化数据使用数据抽取（Information Extraction）和处理非结构化数据的自然语言处理。

统计分析

统计分析是复杂数据分析的常规武器，运用统计方法进行定性和定量的分析，对研究对象产生更为深刻的认识。主要的分析技术包括：

- 假设检验（Hypothesis Test）来推断假设是否成立
- 显著性检验（Significance Test）来检验变量对目标的影响程度
- 相关分析（Correlation Analysis）来分析变量之间的关系
- T 检验（T Test）来比较两组数据是否存在显著差异
- 其他统计分析方法，如方差分析（ANOVA）等。

数据建模

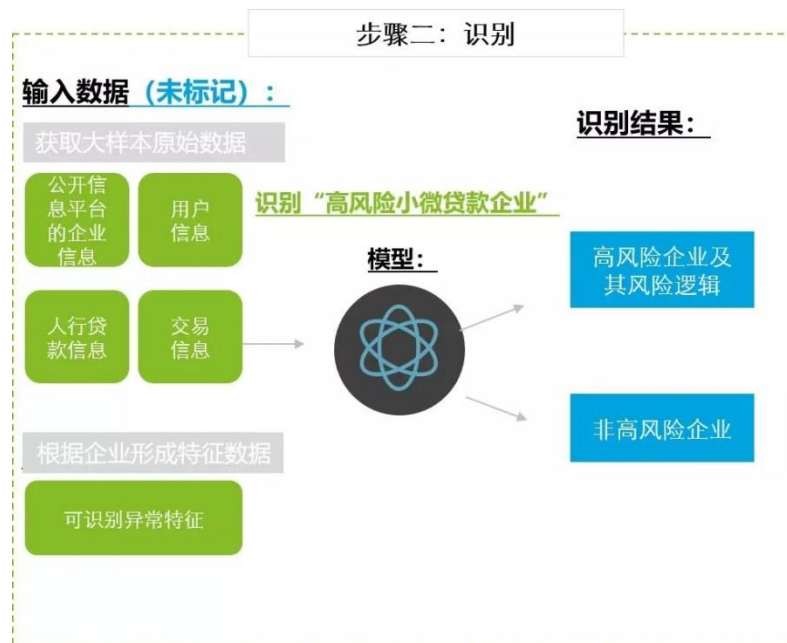
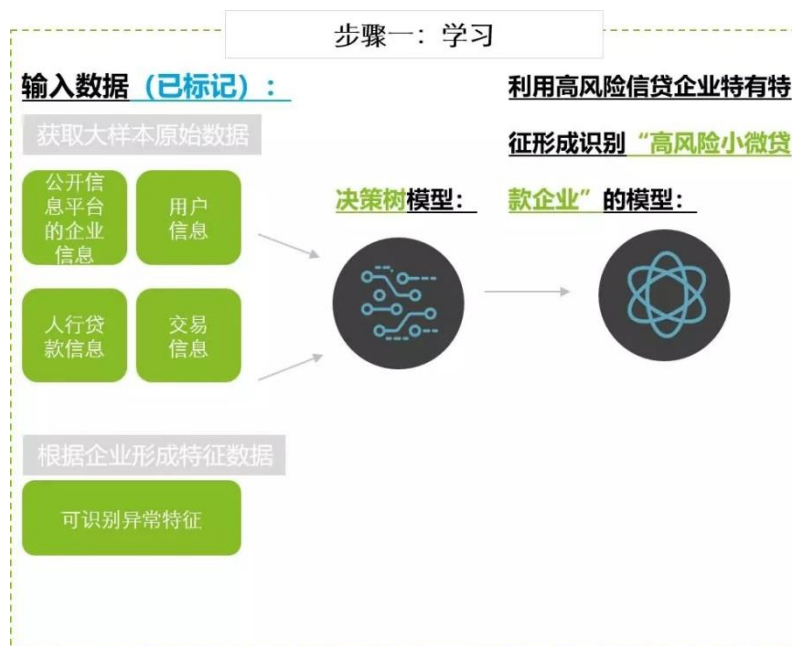
数据建模是复杂数据分析的核心技术，能从大量数据中通过算法搜索隐藏于其中的信息。主要的算法包括 分类（Classification）、聚类（Clustering）、预测（Prediction）、估计（Estimation）和复杂数据类型挖掘（text、Web、图形图像、视频、音频等）等，数据建模可基于用户行为数据实现对人群特征的分类、根据历史数据预测行业走势和基于对海量的图片深度学习做到对图形图像的识别。

3、实战

近年来，银行的小额贷款业务发展迅猛，无论是小微企业贷款还是个人小额贷款都呈现出井喷式增长。面对快速增长的业务量，银行内部审计面临不小的压力，若按传统贷款审查流程对小额贷款进行全面检查，内部审计工作量不可承受且效益低下，若采用抽样审计，往往又容易以偏概全，存在较高的检查风险。

某概念验证 (Proof of Concept) 案例

为了对小额贷款业务进行有效的审查，某银行内审部引入了复杂数据分析手段来应对 大量的小额信贷审计工作。通过搭建统一、共享的内部审计数据平台，广泛采集客户的各类数据，将银行内部生产环境中获取的小额贷款相关的“一手”业务信息和网络爬虫在互联网上爬取到的客户个体和组织相关信息进行交叉对比核实，形成完整真实的客户信息和行为。通过对用户信息的分析和挖掘，对历史逾期、账户资金流动、个人征信、客户信用风险等级等特征打上标记。将整理好的数据切分为训练数据集和测试数据集。在训练数据中，把发生过违约的人群作为目标建立机器学习模型，分析每个用户特征对判断客户信用资质和信用风险高低的影响，找出高风险企业的特征，持续调整优化模型，提高模型准确率，并使用测试数据验证模型。通过使用机器学习模型发现小微信贷中隐藏的规律、规则和趋势，从而达到总结过去预测未来的效果，给小微信贷业务的风险防控提供有价值的内部审计建议。



4、奥义

内审部通过引入复杂数据分析并与德勤的智能化一揽子方案相结合，实现了内部审计的从传统审计到智慧审计的飞跃。以复杂数据分析为导向的智慧审计降低了内部审计成本，提高了内部审计效率。在充分将银行自有数据资产变现的同时，引入外部大数据，生成一体化的解决方案，帮助银行内部审计在面对银行的新业务、新风险、新

环境核心挑战时拥有更强的洞察发现力以发挥更大作用。

三、结语

随着近年来我国经济下行压力加大，小额信贷风险进一步提高，银行内部审计作为第三道防线，越来越需要练就一双“火眼金睛”，而复杂数据分析就像是一副眼镜，可以提高内部审计的“分辨率”，提炼数据价值，帮助实现银行内部审计工作在持续性、立体性、全面性和智能性四个方面逐步升级。

(原 文 链 接 :
<https://mp.weixin.qq.com/s/bPa14P32Yhmy0kDfQt1Plw> , 转载请注明。)