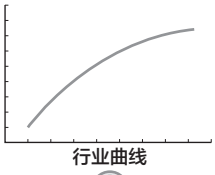
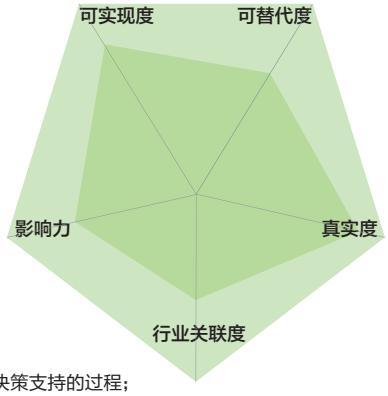
DOI：10.3969/j.issn.1001-8972.2018.05.021



link appraisement industry

王威巍
中国银行卢森堡分行

point



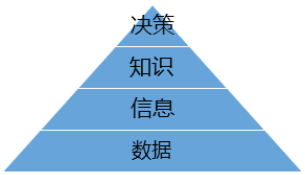
本文主要介绍了数据挖掘技术在 Fin – tech 领域的应用，阐述了数据转化为信息，知识并最终提供决策支持的过程；数据挖掘常用的算法及其适用的场景；数据挖掘与模型搭建的关系，以及如何搭建一个行之有效的数据模型。

Fin – Tech 与数据挖掘

关于数据挖掘

作为计算机科学经典的分支，数据挖掘结合了计算机科学，AI（人工智能），机器学习，统计学等多种学科，从大型的数据集合中挖掘未知的，可理解，可执行的信息并用它进行关键商业决策的过程。

如果我们把大数据比作海洋，我们想要获取的信息为鱼，而数据挖掘就是捕鱼的网。可以说，数据挖掘是驾驭大数据，实现大数据价值的重要手段。由于大数据是大体量，来源丰富，结构复杂的数据源，用传统的方法往往很难充分利用大数据，为了能够有效利用大数据，则需对其进行收集，存储，萃取，转化，加载，分析。下图简单描述了数据是如何转化为决策的。



该金字塔结构的最下层为数据，它是由存储在各类型数据库或数据仓库中结构化，非结构化以及半结构化的数据源构成；数据通过清洗，萃取，转化后得到了信息，它是数据的具体表现形式；对信息进行业务处理后，得到的可用于生产，有意义的信息被称之为知识。最终这些知识将对领导层的重要决策进行佐证和支持。数据挖掘技术就是这个过程的引擎。我们可以说是数据挖掘成就了大数据演绎：“数据转化为信息，信息提炼为知识，知识引导决策”。

数据挖掘的主要内容包括建模和算法，在具体应用过程中就体现为：表述问题和解决问题。

表述问题的本质是把实际的业务问题转化为数据挖掘问题，也就是把客观业务标准化，数据化。

统计学将一些结构化的数据称之为“Hard data”，这些数据具备清晰的数据结构和数据逻辑，与各类数据库及分析软件具有良好的兼容性，经过常规的数据处理即可进行分析处理。而现实生活中往往存在的是一些非结构化的数据，被称为“soft – data”：例如一些音频，视频，图像，图表之类的数据。将业务问题转化为数据挖掘问题就是将业务相关的“soft data”转化为机器识别的“hard data”。这个过程我们可以借助经济学中的“经济人”假设，或者采用效用函数（utility function）等模型来实现。

解决问题就是根据实际的业务需求，采用合适的算法建模，通过数据挖掘的方法解决该问题。

当我们把业务领域的问题清晰的转换为可量化处理的数据模型后，即可根据实际情况，采用不同的算法模型，对不同的数据模型进行数据挖掘处理，从而获取到数据背后的关联和依赖关系亦或者预测出未来的变化情况和发展趋势。

数据挖掘常用算法

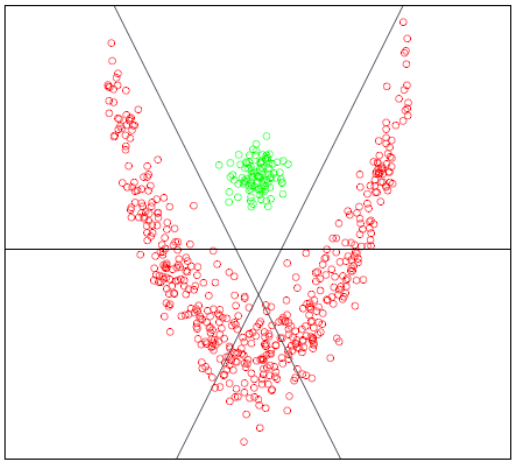
算法可以看作是数据挖掘的灵魂，对应不同的业务场景因地制宜，采用合适的数据挖掘算法往往可以事半功倍。例如对客户进行风险评估时采用分类算法；预测金融市场波动时采用回归线算法；风险管理时采用的关联性算法等等。在金融行业较为常用的数据挖掘主要包括如下几种。

分类算法：

分类算法通过对已知训练集进行分析处理，从中发现分类规则，以此来预测新数据的类别。分类算法在金融领域的主要用于风险评估，客户类别分类，异常检测等。例如通过整合数据库中的客户信息，构造一个客户风险的模型，从而对新产生的客户贷款请求进行风险评估。

常用的分类算法包括：Bayes，Decision Tree，SVM，SNN，神经网络等，欧盟著名的反洗钱系统 SIRON AML/KYC and Embargo 系统以及美国的 FAIS

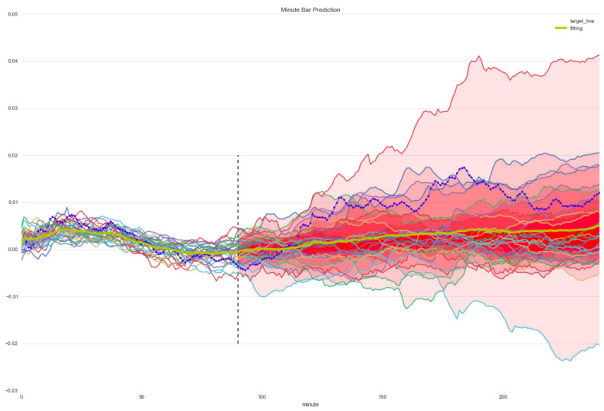
AML 系统其分析模型的核心就是分类算法。伴随着科学技术的飞速发展，AI 学习和深层网络学习都将更好的服务于反欺诈和风控方面。



美国主流的个人信用评级工具 FICO 其主要的算法逻辑便是：获取申请人的全量资料，然后将申请人的资料和数据仓库中所有贷款人的资料进行比较，其中包括日常消费情况，透支情况等，同时结合聚类算法和回归算法对申请者进行风险评级。

回归分析

回归分析是研究自变量和因变量之间的关系，建立两者之间的回归方程，然后基于此回归方程构建预测模型，随即便可根据自变量在预测期及预测范围内的变化来预测因变量的一种预测性算法。通过这种算法可以预测很多领域中各个因素（key Indicator）之间的关系。例如：预测股票，金融市场波动情况预测，产品满意度预测等。其主要的算法包括：线性回归，逻辑回归，多项式回归等。



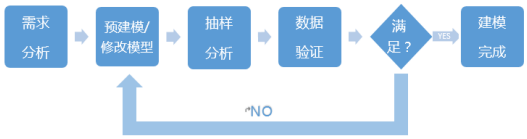
关联性分析

关联性分析是一种发现大型数据库中某些变量之间存在相关性（已知的，更多情况下是强调发现变量之间客观存在的，潜在的关联性）的算法。它的作用是利用一些自定义的量度来之别数据库中的强规则。可分为简单关联，时序关联，因果关联等，除了被人津津乐道的“尿布啤酒组合”，该算法越来越多的被应用到金融领域。例如通过对客户的违约行为进行关联性分析，挖掘出与违约客户有强关联的变量。通过着重对这些之前未曾注意的变量的监控，从而加强了企业的风险管理；亦或者对 KPI 数据进行关联性分析，找出数据中的异常值，进而挖掘出其关联或者非关联数据的异常部分。

数据挖掘与模型搭建

数据挖掘的核心同时也是其最重要的就是将业务问题抽象转化为数学模型。在日常的业务中，企业会处理大量的数据，盲目的去做数据分析，简单的得出一个数据模型并不困难，问题在于该模型具有多少实际意义？能起到多大作用？

如何确认数学模型的 Input 和 Output 是数据挖掘的关键步骤。这里可以借助机器学习的思路：首先邀请业务骨干或者该领域的专家和 IT 人员一起进行全局的需求分析，建立初步的模型；然后进行抽样分析，让业务骨干或者专家对该模型的结果进行满意度分析；根据专家的意见再次修正模型，反复该步骤直至该模型的分析结果符合准确有效。该过程流程图如下。



目前存在的问题是重算法而轻模型，过度的重视数据处理：采用复杂的算法模型，借助强大的分析软件进行数据处理，而忽视了模型的可靠性和健壮性。事实上，在没有一个优质模型的前提下粗暴的进行数据分析，得到的结论往往都是盲目的，片面的。通过反复的抽样和修正才能真正建立一个行之有效的数学模型。（需注意的是：反复修改模型，并不是从主观上向“标准答案靠拢”，而是从客观上完善该模型，使之可以有效的进行数据的处理和分析，得到客观准确的结果）。

值得注意的是：在建模的过程中，“主观概率”是一个不可避免的问题，客观存在的个人因素或者不确定因素往往会左右模型的 Input 和 Output。“主观概率”会导致每个人（包括专家团队）都会根据自身的喜好不由自主的去屏蔽或者放大一些数据的影响。这就需要通过对模型的反复训练，用数据和结果来论证，以减少个人因素对模型产生的干扰，以确保我们的预建模型有着严谨，清晰的 Input 和 Output。

通常来讲，金融业至少需要搭建以下两个数据模型。

用户模型：用以描述用户特征的模型，描述的内容包括用户经济学属性和社会学属性。经济学属性包括用户的收支情况，资产负债情况，购买力等；社会学属性则包括用户的家庭情况，兴趣爱好等。

系统模型：银行可根据自身发展及合规需求，搭建相应的系统：例如用于进行风险管控的风险识别模型；用于用户信用评估的信用等级评估模型等。

综上所述，一个完善的数据分析体系，一个有意义的数据挖掘项目，首先需要因地制宜，采用合适的数据挖掘算法，其次需要一个良好的理论模型用以指导分析，然后借助实际数据的反复训练逐渐完善该模型。最终便可从浩瀚的数据仓库中挖掘出有价值的“珍宝”，为决策者提供可靠的决策支持。

李彩霞 P95

《中国科技信息》杂志微智库成员

个人简介

李彩霞，年龄，33，性别，女，最高学历，研究生，学科专业方向，情报研究、专利分析，职称，助理研究员，学位，硕士，所在单位，吉林省科学技术信息研究所，微信，licaixia-0207，Email，caixia.ice@163.com，手机，18643062756。

教育背景

2000—2003：吉林市第四中学
2003—2007：吉林农业大学
2007—2010：吉林农业大学

工作经历

2010—至今：吉林省科学技术信息研究所

建议观点

1、科学技术是第一生产力。首先要完善领导的科技责任制，确立考核制度，实抓生产力，而且要坚持不懈的抓、一任接着一任的抓，要让领导认识只有科技进步、只有自主创新，才能发展中国。
2、提高企业的自主创新能力。充分发挥企业的科技自主创新地位，要让企业意识到，如果不抓科技进步不抓自主创新，企业就寸步难行，没有发展前景，要鼓励和支持企业自主创新，打造不可复制的核心竞争力。
3、加快科技人才的培养和引进。为科技人才提供再学习和进修的机会，将所学的知识结合自身的工作，勇于创新、大胆创新。强化各行各业科技人才队伍建设，使科技人才永葆生命力。

王威巍 P60

《中国科技信息》杂志微智库成员

个人简介

王威巍，年龄，30，性别，男，最高学历，硕士研究生，学科专业方向，计算机科学，商业智能，职务，计算机工程师，所在单位，中国银行，微信，weiweiwang2015，Email，Weiweiwang8077@gmail.com，手机，+352 661556734，兴趣爱好，旅游，运动，掌握语言，汉语，英语，法语。

教育背景

2003—2006：陕西榆林市第一中学
2006—2010：西安邮电大学 -计算机系
2010—2014：法国洛林大学，计算机商业智能
2016—2017：卢森堡大学，MBA

工作经历

2013—2013：Business Decision， BI 工程师
2014—2015：中国工商银行金融租赁，BI 工程师
2015—至今：中国银行卢森堡分行，网络工程师

建议观点

1. 伴随着银行业的海外业务的快速扩张，日益“严苛”的当地监管要求，使得银行也在谋求快速发展的同时，也必须加强自身的法律合规意识，如何有效利用现有的数据，结合信息科学技术实现高效，准确，及时的数据分析，充分保证监管要求成为中资海外银行需尽快解决的问题。
2. 银行业自身拥有大量的客户和交易数据，如何有效利用这些数据，挖掘出数据中隐藏的财富，建立有效的风险评估机制；进行精准的客户营销以及使用数据挖掘技术提供决策支持，成为银行业日益关注的问题。
3. 大数据，数据挖掘，Fin-tech 以及 AI 技术的发展，给银行业带来了新的机遇和挑战，如何将银行业和 IT 技术充分结合，使之相互作用并相互促进，是银行业在互联网时代面临的最重要的问题。