

# 数据挖掘在金融欺诈检测和预防中的应用

期刊：金融电子化 2010 年 第 8 期      作者：肖可砾 熊辉

---

数据挖掘技术是对数据进行深层加工和分析的信息处理技术。作为当今计算机科学中的一个新兴领域，其应用遍及金融、贸易、服务、医疗、教育、科研、军事等各个领域。其中，金融欺诈行为的检测和预防是其在金融领域的新应用。在庞大的金融市场中，每个交易者都在努力为自己创造价值。交易者掌握信息越充分就越有可能在交易中获利。因此，现代金融市场安全的研究必须引入行为学和信息学的概念。人们一直在尝试用各种方法来减少金融服务的不安全性。及时发现影响安全的行为至关重要，而面对涉及人类金融活动行为和信息传递这种大数据的处理和分析，数据挖掘技术的使用就显得十分重要了。通过数据挖掘技术，对客户及员工行为以及金融交易双方的行为进行归类并对其进行监督，可以有效跟踪检测违规操作等欺诈行为并可以对其可能带来的损失进行预测，从而达到防范金融风险的目的，甚至可以预防金融危机的产生。

## 一、金融欺诈的行为特征

金融欺诈作为当前社会广泛关注的一个课题，其定义也是多样的。一种说法是，金融欺诈就是利用金融产品规则上的漏洞（如银行业务流程中的漏洞或金融监管中的漏洞）获取利益的违法行为。也可以说，凡是在金融市场以谋取自身利益为目的，对他人或组织造成利益损失的违法行为，都是金融欺诈行为。因此，面对金融市场中万千的金融产品以及它们的衍生物，金融欺诈行为自然是多样化了。

金融欺诈行为可以以种形式出现，按涉及的金融产品分有：贷款欺诈、存款欺诈、票据欺诈、证券欺诈、银行卡欺诈、保险欺诈和金融衍生产品欺诈等；从行为的来源分，可以分为外部威胁和内部威胁两类；按欺诈手段，可以分为三种类型：

**一是利用银行交易系统，进行非法侵入或违规操作，谋取不正当利益。**信用卡欺诈、身份窃取以及大量的内部违规操作等属于这种类型。最典型的有业务流程欺诈，即利用业务流程上的漏洞获取利益的欺诈行为。首先，从银行业务来看，常见的表现形式有恶意挂失银行卡和恶意透支等。这类欺诈大多来自银行内部。因为只有柜员才有可能掌握大量的用户资料，包括用户姓名、账号、身份证复印件等。并且他们对每一项柜台业务都十分熟悉，从而有可能从业务流程中找到漏洞。例如，柜员可以通过客户复印件挂失银行卡并重新设置密码。同样，柜员也很容易掌握信用卡账号和客户的基本信息，自然就可以进行恶意透支。另外，从投资业务来看，越权交易是可能给公司带来极大的损失。投资银行一般都会更具其可能承受的风险来设定交易员的最大交易次数和交易额度。如果一旦交易员通过交易流程中的某些漏洞，越权交易，则可能给公司带来巨大损失。如，2008年1月，法国兴业银行的一个期权交易员 Jerome Kerviel，利用交易系统漏洞越权交易，给兴业银行造成 49 亿欧元的历史最大损失。2005 年，中国航油（新加坡）股份有限公司原新政总裁陈久霖因违规炒卖新加坡期货指数导致公司 5.54 亿美元的巨额损失。由于基于业务流程的欺诈行为大多在银行或公司的交易数据库中可以找到信息，所以一旦发现问题，对执行者的定位是很迅速的。更进一步，通过计算机技术我们能提前锁定那些潜在的欺诈行为。如一段时间内某柜员多次办理挂失手续，或某交易员超出权限交易等。而对于金融交易这类大数据库，我们应该用到数据挖掘的技术来进行分析和处理。

**二是提供虚假承诺或虚假信用保证资料进行欺诈。**大量的投融资欺诈属于这种类型。欺诈者常以高利润投资为诱饵，不断获取投资者输入的一种金融欺诈手段。投资欺诈的发起者往往是一些知名的人士或企业。他们宣称自己即将展开的投资计划，并承诺极高的回报率（年回报率通常大于 20%）。事实上，投资计划并没有进行，而那些高利率回报则是由后注入的资金来支付的。最终，大部分投资者都血本无归。这种非法集资的手段也称为“庞氏骗局”。“庞氏骗局”是以查尔斯·庞兹（Charles Ponzi）命名的。1903 年，庞兹从意大利移民美国。从 1919 年开始，他策划实施了一个投资骗局。他许诺投资者将在三个月内获得 40% 的投资回报。其实最初投资者的回报都是有新投资者的资金来支付。他成功在七个月内吸引了三万名投资者并持续一年多后才被人识破。近年来最有名的一个投资欺诈就是麦道夫的“庞氏骗局”。其受害者包括奥地利银行、瑞士银行、汇丰银行、通用对冲基金 Tremont 公司等知名银行和基金。麦道夫的庞氏骗局其实早在 1990 年就已经开始，但是直到 2008 年金融危机才被发现。这其中四个主要的原因。第一，麦道夫利用奢华场所建立人脉关系，并接触投资者。这些投资者往往对自己的投资不太关注。第二，他早期树立了投资必有回报的口碑。第三，通过朋友、家人和生意伙伴发展新投资者。第四，设置较为合理的投资回报率（年回报率 12-13%）。这些设计让麦道夫的投资骗局很难被揭穿。甚至在他被捕前，他依然向公司员工发放剩余资金。当然，通过对一些重要特征的分析，投资欺诈也是可以预测的。其一，回报率过高。高回报率可能吸引投资者不断加入，从而维持资金链。但事实上，资金却没有在市场上产生价值。其二，过于复杂的投资计划。金融实际上比想象中要简单直接。任何一个金融投资计划都应该可以被清楚地解释。所以，如果投资计划被渲染得过于复杂，这样的投资就必须小心。其三，只有极少人监督资金运作过程。麦道夫的例子告诉我们，他独自管理投资

者的资金，这样几乎没有人能知道资金的流向。这样，如果我们对投资计划的以上三个方面进行评估，就能初步预测并锁定投资欺诈行为。

票据欺诈也是一种常见的这种类型的金融欺诈。票据融资是企业融资的重要渠道之一。同时，也是商业银行拓展业务的重要方向。统计数字表明，从 2000 年到 2006 年，中国的票据融资余额年增长 70%。票据融资给银行和企业都带来利益。因此，票据融资中的违规操作而引发的重大案件让人们开始关注票据欺诈的形式和过程。目前，广泛使用的票据有支票和汇票。汇票包括银行汇票和商业汇票，其中商业汇票根据承兑人的不同又分为银行承兑汇票和商业承兑汇票。首先，从银行角度来看，如果盲目追求业务量，而对企业提供的商品交易合同等文件未经严格审查，这样就可能为其签发无贸易背景的银行承兑汇票。而由于企业无力支付到期票据，银行又为了掩盖不良资产，只能继续为其签发银行承兑汇票，导致恶性循环。商业银行与企业联手包装票据也是一个严重的问题。为了逃避监管和获取利息收入，银行将票据背书转让给信用较好的企业，并由它提供其实与汇票无关的商品交易合同，增值税发票等。到银行办理贴现后，再将资金转回真正需要贴现的企业。在争抢票源的恶性竞争中，有些银行严重违背票据融资的宗旨。如，设置低于人民银行再贴现率的贴现率，或将资金转入个人账户等。另外，从企业的角度来说，一些企业为了用票据办理融资，提供虚假商业合同，财务报表等。大多这种情况下，企业往往不能按时还贷。

**三是通过隐瞒重要信息，人为制造信息不对称进行欺诈。**在证券市场大量的内幕交易、在衍生产品推销过程中故意隐藏可能产生的风险性、运用各种手段操控市场以期套利等，都属于这种类型。由于满足各种投融资需求而形成的各类证券、满足分散和配置风险而创设的各种衍生产品，其收益和风险的决策具有复杂性、可变性和主观性。而且这类产品的供求常不具备充分竞争市场的特征。为

金融机构制造信息垄断提供了先天条件。使这类金融欺诈具有很大的隐蔽性。这些领域又先天是监管的薄弱环节，常可大行其道。而由于其覆盖面广和价值链长，会产生危害极大。2007年，由美国次级抵押贷款市场所引发的次贷危机，以其衍生的抵押化债务债券（CDO）、信用违约掉期（CDS）等次贷证券化产品为传递链向全世界扩散。在这些以指数级放大的传递链下，全球金融危机爆发。这其中大量利用信息不对称进行的欺诈行为。各国在合作对抗危机同时，对金融市场的安全和稳定也越来越重视。很多人开始注意到这次金融危机的根源，其实可以归结为金融欺诈行为。最近披露的高盛公司的欺诈丑闻，正是对金融危机根源的深刻反省。

## 二、数据挖掘技术应用于金融欺诈的检测

欺诈所涉及的交易行为，一般具有非正常或非公平交易的属性，由于缺乏公平公正的交易动机和与实体经济活动相一致的资金运动规律，或有异于一般客户和账户运用的行为特征。从而呈现出各种异常的特征，包括交易所行为异常、交易对象异常、交易数量异常、资金走向异常等。如果能建立一定的数量模型，发掘和识别异常信息，并及时发现问题，避免损失。数据挖掘技术正是在这一领域有广泛应用前景。

所谓数据挖掘，就是运用自动或半自动计算技术从海量数据中发现隐藏的、新颖的、具有潜在价值的信息或有意义的模式帮助决策支持。作为一门交叉科学，数据挖掘综合了统计、机器发现、模式识别、数据库、信息理论和人工智能等多学科的先进技术和想法。通过数据挖掘技术，我们可以从海量的金融交易数据中找到用户欺诈性行为。比如发现交易环节市场操纵行为。

数据是数据挖掘的基础，在对数据进行分析之前，我们必须对数据有足够的

了解。所谓数据就是在计算机中处理过或准备处理的任何形式的数字、文字或图形。从数据的表现形式来看,我们可以把数据分为定性数据和定量数据。定性数据的特点是我们可比较数据的大小、相等或不等,却无法或没有对其精确量化。例如,邮政编码、身份证号、信用评级。另外,定量数据即那些可以被精确量化的数据。例如日期、交易量、货币量、年龄、长度等。从记录对象来分,我们又可以把数据分为事务数据、非事务数据和元数据。举一个简单的例子,对一个销售公司来说,事务数据就是与公司运营直接相关的数据,包括公司销售量、成本、库存、员工工资表、财务信息等。非事务数据则是那些公司外部与简介相关的数据,包括市场销售量,预测经济走势等。另外,元数据则是指那些标记数据库逻辑的数据。

有了数据,就有了数据挖掘的可能。即识别大量原始数据中有利于我们分析问题的特征数据的可能。目前,最常见的数据挖掘技术包括分类预测( Predictive Modeling )、聚类分析( Cluster Analysis )、关联分析( Association Analysis )以及异常诊断( Anomaly Detection )。

分类预测主要是使用历史数据建立分类预测模型,并用所建立的模型对未来数据进行分类预测。分类预测的方法包括树型结构的分类(Tree-based)、基于规则的分类( Rule-based )、最近邻居法( Nearest Neighbor )、递归法( Regression )、人工神经网络法( Artificial Neural Networks )、绘图法( Graphical Methods )、以及向量机 Support Vector Machines ( SVMs )。这些方法可以用于解决离散型数据和连续型数据分类预测问题。如果能通过对非法集资、洗钱等典型诈骗行为的逻辑路径分析找到其行为特征就可利用上述方法挖掘出其相关数据,以检测出其诈骗行为。

聚类分析的职能是把一个数据集的所有数据点分到不同的组里,从而使属于

同一组的数据更相似而不同组的数据有很大差异。聚类分析的方法也包括 K - 均值聚类 ( K-Means )、自组映射 ( Self-organized Maps )、高斯混合模型 ( Gaussian Mixture Models )、分层聚类 ( Hierarchical Clustering )、子空间聚类 ( Subspace Clustering )、图形算法 ( Graph-based Algorithms ) 以及基于密度的算法 ( Density-based Algorithms ) 等。这些技术的区别在于用不同的方法计算数据点之间的距离以及定义聚类成员。聚类分析在商务智能及决策分析领域获得了广泛的应用。例如,应用聚类分析的方法,我们可以发现拥有相似价格的运动模式的股票聚类。从中可能发现关联交易及内幕交易的可疑信息。

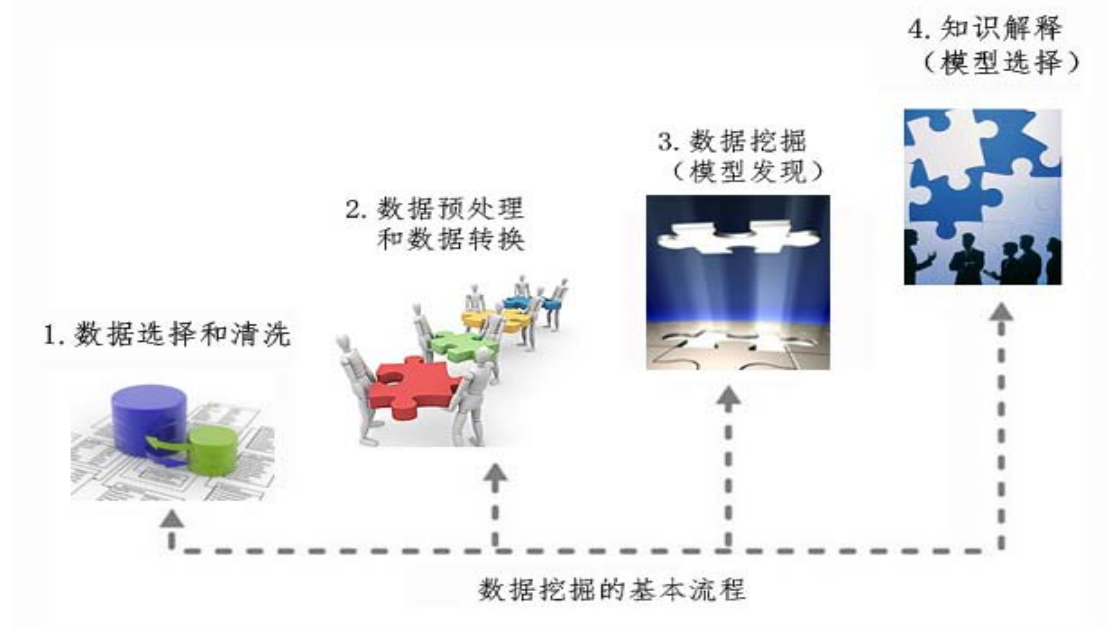
关联分析的主要作用是从海量数据中高效、准确地发现强相关性事件。关联分析的方法包括:关联规则分析 ( Association Rule Mining ) 和统计相关计算 ( Statistical Correlation Computing )。关联分析对于发现隐藏在数据中事物的内在联系有很大的作用。举一个例子,关联分析可以用于监控多个用户的关联交易行为。这就给检测跨账户协同进行的金融欺诈提供了有效的路径。即使欺诈者的交易行为表面上看起来属于正常交易行为,我们还是可以通过关联分析进行跨账户协同检测来找到其中的特殊状况。

异常诊断也被称之为偏差检测。其主要目的是搜寻并发现数据中的异常点或异常事件。一种常见的方式做异常检测是构建一个数据的正常行为档案,并用它来计算其他观测对象的异常指数。直接可以挖掘识别规则范围外的行为或其它异常行为。如今被广泛使用的异常诊断技术包括基于统计原理 ( statistical-based ), 基于距离 ( density-based ), 和基于聚类分析 ( cluster-based ) 的异常诊断技术。试算出,财务指标之间的相关关系的正常分布后,也很容易对明显背离的公布的数据作为异常信息检测出来。这些技术对于金融风险管理,客户信息安全管理,和网络安全管理等领域到头重要。通过这

些技术,我们还可以发现公司提供的财务报表中经常存在异常夸大的收益和虚报的收入等可疑情况。例如,通过分析安然公司的财务报表,我们本可以提前诊断和预防金融欺诈。首先,该公司应收账款相对于销售额的增长。判断这一增长是否合理的一种常用的方法是与同行业的平均值做比较。这样,我们可以较有效的判断该数据异常是否意味着公司的潜在问题。其次,存货周转率递减的趋势也可能意味着公司存在不合理的仓库管理。我们可以通过对比同行业的平均水平来判断该参数是否合理。第三,公司的现金流与净收入之差应该保持在一个合理的范围。如果现金流远远低于净收入,很有可能该公司虚报了净收入。此外,主要业务人员的关联信息也很重要。分析人员在安然公司财务报表的页脚中发现安然的资深员工安迪·法斯托(Andy Fastow)在担任安然公司财务总监的同时也是另一家与安然公司有交易行为的经济实体的主要合伙人。安迪·法斯托利用自己安然公司财务总监的身份从安然公司窃取了大量资金。

### 三、金融欺诈数据挖掘的基本流程

完整的数据挖掘的基本流程包括数据选择和清洗,数据预处理和转换,数据发掘及模型发现,以及知识解释等四个步骤。对于金融欺诈的数据挖掘也有相同的流程。



## 1) 数据选择和清洗

对于应用数据挖掘进行金融欺诈检测，首先要考虑哪些数据是有用的，可以从哪里获得这些数据？这是数据挖掘的第一步。2008年注册软件欺诈检验国际组织 (Association of Certified Fraud Examiners) 提供的职业欺诈报告中统计，有55%职业欺诈来自于管理失误，30%来自员工，只有大约6%来自有组织的犯罪行为，5%来自客户。可见大部分的欺诈行为来自公司内部。因此公司财务报表应该作为数据的一个重要来源。公司财务报表包括资产损益表，收入表，现金流量表以及预算汇总表，销售和服务表，交易列表等。这些金融报表可用于分析公司的财务状况，从中找出会计操作等财务违规行为。此外，公司内部电子邮件网络提供的电子邮件来往数据和公司内部系统提供的员工工作状态信息可以用于检测员工内部操作等违规行为。同样，股票和期货的交易数据，商业和经济网络提供的商业来往数据也可以用于检测市场操作等违规行为。银行异常交易信息的检测，须从银行交易系统获取数据，有转发方式、网点终端操作层及网络层等不同方式获取检测数据。

## **2) 数据预处理和数据转换**

数据挖掘通常需要处理海量的原始数据。数据预处理和数据转换的过程则是对原始数据进行加工。该过程包括数据清洗，数据整合，特征向量提取等众多辅助数据产生的方法。目的是为了提炼出可以作为欺诈检验参数的特征数据。例如，员工工作情绪可以用员工工作时间变化，休假情况，收入/消费水平等数据来设计检测模型。此外，员工的性别，教育程度，犯罪记录，在公司工作时间等数据可以用于设计欺诈检验模型。

## **3) 数据挖掘/模型发现**

这一过程，是通过运用数据挖掘的四大类技术：分类预测（ Predictive Modeling ），聚类分析（ Cluster Analysis ），对收集的数据进行实时的欺诈检测，对欺诈的潜在环节或个人进行定位，并找出隐藏欺诈模式。例如，环交易，用户重复支付，重复发票，重复挂失，异常大额消费或存款，员工业务量异常变化，非法授权过程等可以迅速被确定并生成报告。对于严重违规操作可以通过系统进行紧急处理，如取消员工交易权限，冻结账号等。

## **4) 知识解释/模型确认**

对生成的挖掘报告，可以做专家人工核对，对结果进一步确认。剔除误判。并对系统模型进行调整改进，对于金融欺诈诊断来说，这一步就是确认欺诈行为的一个重要步骤。数据挖掘是一种自动或半自动的筛选技术，可以从海量数据中发现大量的可疑事件，而只有对其做进一步专家确认，才能提高挖掘的有效性。当然，我们通过这一步骤，除了可以确认真正的欺诈行为，同时也确认了那些被误判的行为。这样，通过对误判行为的特性分析，我们就可以建立更精确的数据挖掘方法，提高数据挖掘的效率，从而减少人工确认的工作量。

最后要说明的是以上四个步骤是个循环反复的动态过程，只有在动态运行过

程中，我们才有可能对数据，挖掘模式进行动态调整，从而有可能把握住不断变化的金融欺诈模式。

通过使用数据挖掘的技术，我们可以动态调整欺诈诊断模型，对各种金融数据进行快速扫描并发现潜在欺诈风险。那么，各类金融欺诈案件的预警，确认速度和准确性将会得到大幅度提升。目前数据挖掘技术在金融机构案件防范中正在获得广泛运用。中国建行、农行等甚至相当多的中小银行都建立了事中的监督系统，及时识别异常交易信息和操作信息。通过这些检测系统，实时对交易数据进行获取和分析，对于符合预警规则的敏感交易进行风险预警提示并记录；根据各类业务要求对敏感交易，如高频、高额、可疑和异动等业务设置预警规则，并根据实际情况对前台网点的大额、可疑、高频、特殊等敏感交易进行检查记录；对储蓄、会计、信贷、信用卡、ATM/POS、网银、大额支付、冲正流水、挂失解挂、内部账务等的全部流水数据进行监测。在第一时间发现和防范操作风险，减少损失。央行就反洗钱等建立了动态识别系统。证监部门、财政部门就企业财务合规性也运用了专门的检测模型。但针对需要处理大量信息的金融风险监测问题，仍是一个需要不断深化研究的问题。如运用神经网络和决策树技术等挖掘数据信息，尽早发现各类欺诈行为的预警，对内部交易及人为制造信息不对称的上文第三种类型的欺诈的识别分析尚未建立十分有效方法。日益进步的数据挖掘技术在解决这些问题上的应用也将会给金融安全技术的发展带来新的手段和动力。

(肖可砾，美国 Rutgers 大学金融学博士生；  
熊辉，美国 Rutgers 大学商学院终身教授。)