文章编号: 1003- 207(2003) 01- 0038- 04

一种基于距离的欺诈风险分析方法

柳炳祥[1], 盛昭瀚[2]

(1. 东南大学经济管理学院, 南京 210096;
2. 南京大学管理科学与工程研究院, 南京 210093)

摘 要: 叙述了欺诈的基本概念, 分析了欺诈风险产生的原因, 研究了欺诈风险的识别、分析和评价方法, 指出了传统的欺诈风险分析模型存在的局限性, 提出了一种基于距离的欺诈风险分析方法, 并通过一个信用卡欺诈的模拟实验说明了该方法的可行性和有效性, 为欺诈风险的分析提供了一种新的思路和分析方法。

关键词: 欺诈; 距离; 意外规则; 离群数据; k 个最近邻居点
中图分类号: F224.3 文献标识码: A

1 引言

随着信息技术的迅速发展和网络化经济的快速进步, 传统的商业模式发生了根本性的变化, 独特的竞争优势难以获得, 经营比以前更具挑战性, 成功的企业利用其强大的技术力量和无形资产创造财富的同时面临的风险也越来越多, 越来越复杂。如何对各种风险进行识别、分析、评价和管理, 是企业取得竞争优势, 实现可持续发展战略的关键。因为风险总是与机遇并存, 只有全面了解并有能力控制其面临的风险, 企业才能更有效地寻求机遇, 获得发展。在企业面临的所有风险中, 有一类风险即欺诈风险变得越来越普遍, 危害也越来越大。一旦发生欺诈, 企业将面临管理活动的失败、市场份额的丧失和商务活动的失败, 导致企业失去业务、顾客、竞争力和信誉。据统计, 在英国, 62% 的企业认为欺诈行为比以前更加普遍, 三分之二以上的企业曾经发生过欺诈事件, 这还不包括那些发生了欺诈行为但由于各种原因没有向外界以布的企业[1]。由此可见, 企业间的欺诈行为是非常普遍的, 而且一旦发生, 给企业带来的损失是巨大的。如何准确、及时、有效地预测到企业可能发生的欺诈行为是非常有意义的。本文利用数据挖掘技术中的离群数据的挖掘方法, 提出了一种基于距离的欺诈风险的分析方法。

收稿日期: 2002- 01- 29
基金项目: 国家自然科学基金资助项目(79830010)
作者简介: 柳炳祥(1966-), 男(汉族), 江西九江人, 东南大学经济管理学院在职博士研究生, 副教授, 硕士生导师, 研究方向: 数据挖掘、粗糙集理论、企业危机管理.

2 欺诈风险分析

2.1 欺诈概念

欺诈行为是多种多样的, 很难对它进行定义。在词海中认为欺诈是一种复杂的偷窃形式, 通常以欺骗的方式进行。由于各种不确定性因素的影响和一些竞争对手的存在, 欺诈行为在许多企业中是普遍存在的问题, 如员工的偷窃; 涂改、篡改或销毁数据记录; 做假帐或虚构利润; 兑现假发票、开假赊欠凭证; 恶意透支; 竞争对手利用间谍方式获取商业情报或通过贿赂顾客获得不正当的竞争优势等[2]。

2.2 欺诈产生的原因

要使欺诈行为发生, 需要有多种前提条件, 归纳起来主要有下面四个方面[3]。首先欺诈人一般来说都有一定的动机或原因促使他做出欺诈行为, 如贪婪、占有欲等; 其次要有值得偷窃的资产, 如金钱、产品、原材料、工具、设备等; 再次要有实施欺诈的机会, 如既要有偷窃物品又要有出卖物品的机会; 最后是缺乏必要的监督制度, 欺诈人之所以敢于实施欺诈行为, 是因为他觉得欺诈行为不会被发现, 特别是企业为了减少成本实施业务流程重组, 一般都会减少管理层次, 降低监督的力度, 从而为欺诈创造更大的机会。如果把欺诈产生的原因进行分类的话, 可以分为内部因素和外部因素, 内部因素可控性较强, 是企业自身矛盾的结果, 外部因素几乎难以控制, 是经营环境矛盾的结果, 但欺诈风险往往又是二者综合的结果和体现。

2.3 欺诈风险分析

搜集数据和有关信息是欺诈风险分析的基础。

不断对产生欺诈的因素进行监视, 搜集可能产生欺诈行为的有关数据和资料, 从而及时、准确地做出欺诈风险预测, 并采取针对性地预防措施制止欺诈行为的发生。在进行欺诈风险分析时, 要判断那些资产或资金可能有风险、那些职能部门或人员可能进行欺诈、员工的嗜好和异常情况、发生欺诈的地点和场合、欺诈人通过何种手段窃取这些资产、欺诈的方法、信息系统的安全性以及防止欺诈的监督机制的有效性等。

2.4　欺诈风险评价

对欺诈风险分析之后, 还要对风险进行评价, 以确定欺诈风险发生的可能性、给企业带来的损失及严重程度, 从而为企业制定预防欺诈的措施, 减少欺诈发生的机会, 建立完善的检查监督制度提供参考依据。欺诈风险评价是关于欺诈风险的发生的可能性和损失的危害性计量, 通过对欺诈风险发展趋向进行跟踪和研究, 预先对风险的危害程度及发生的可能性进行估计。

2.4.1　提出分析模型

评价欺诈风险常用的分析模型有头脑风暴法、统计分析法、主观概率法、指标分析法、现场调查法和指数法[4]。头脑风暴法是组织相关的专家共同挖掘可能发生欺诈的所有风险源, 然后将风险进行分类和管理, 并对各种风险进行排序, 以此为依据来处理最严重的一种或一类风险; 统计分析法是对大范围内某种风险发生的历史数据进行统计, 用定量预测的方法对这些数据进行处理, 获取风险发生的概率的方法; 主观概率法是对事物发展的概率分布所作的主观判断, 它不同于根据事物实际发生的次数而统计出来的客观概率, 是一种运用专家的经验, 对某种风险的发生概率进行主观判断, 并以此为基础做出概率预测的方法; 指标分析法是采用各种方法来分析和评价各种定性和定量的风险指标, 为企业提供早期的欺诈风险预警信息; 现场调查法是通过调查企业的场所和环境, 发现所有的风险源, 使管理者明确可能发生欺诈行为的场所和人员, 以便做出相应的预防措施; 指数法是利用风险指数来反映风险对人们的现实威胁的综合性指标, 它是风险危害度和发生概率这两个因素的乘积。

2.4.2　风险评价

识别出各种欺诈风险之后, 下一步就应该评价这些风险。利用上述分析模型对这些风险进行定量计算或定性判断, 并在此基础上得出发生欺诈的关键因素以及对这些风险的评价建议, 以供决策者参考。

2.4.3　已有分析模型存在的问题

欺诈风险产生的原因是非常复杂的, 需要对企业的外部环境、内部经营状况进行分析, 如一个或多个外部环境因素的变化、一个或多个商务进程或系统的不规范或不完善、商务进程和营销活动的不良管理、信息处理过程中的错误、信息系统的不安全性、管理行为不当所引起发的欺诈事件等。这些因素有的是定性的, 有的是定量的, 有的可以度量, 有的却难以度量。由于欺诈风险带有大量不确定因素的半结构化问题或非结构化问题, 产生风险的因素复杂, 种类繁多, 有的因素由于没有历史数据和相应的统计资料, 有的因素和欺诈人的心理、生理及教育程度有关, 很难科学地计算和评估, 只能采用定性和定量相结合的方法确定。运用上述分析模型对欺诈风险进行评价, 存在很多的局限性, 因此需要应用其它技术和分析方法来评价风险。本文采用数据挖掘技术中的离群数据的挖掘方法, 提出了一种基于距离的欺诈风险的预测方法, 借助企业的内部网、外部网和互联网搜集有关的数据和资料, 迅速捕捉到企业可能发生欺诈风险的一切可能事件和先兆, 并对此进行监视、评价和管理。

3　基于距离的欺诈风险预测方法

离群数据的挖掘是数据挖掘方法之一, 所谓数据挖掘是指从数据库的大量数据中揭示出隐含的、先前未知的并有潜在价值的信息的非平凡过程[5]。传统的数据挖掘技术都是针对关联规则而设计的, 支持度和置信度大, 常用于预测, 而且这些规则也是领域专家能够理解的规则, 而意外规则虽然也有高的置信度, 但由于支持度小, 因而常常被忽略, 同时也是经常出乎专家意外的规则。在欺诈风险分析中, 我们更感兴趣的是挖掘出那些意外规则[6~8]。

3.1　基于距离的离群数据的挖掘方法

定义 1: 离群数据: 对于数据集 $T = \{t_1, t_2, \cdots, t_n\}$, U 为数据对象, 如果数据集中有 p 部分数据 S 远离于对象 U, $S \in T$, $s \in S$, 则称数据 U 为离群数据。

定义 2: 距离: 数据集中的 p 部分数据 S 与离群数据 U 之间的距离表示为 $\mathrm{DIS}(p, d)$, 其中 d 是参数, 表示以离群数据 U 为中心的圆的半径, 称 U 为基于距离的离群数据。

定义 3: 离群置信度: 设数据集 $T = \{t_1, t_2, \cdots, t_n\}$, 其中 $c\%$ 的数据为离群数据, 称 c 为离群置信

度, 且: $c = k/n$, 这里 k 为离群数据的个数, n 为数据集 T 中全部点的个数。

定理 1: 数据集 T, n 为数据对象的个数, 对象 U 为离群数据, 其以 d 为邻域包含数据对象 f, $f \in T$ 且 $DIS(U, f) \leqslant d$, $DIS(U, f)$ 是距离函数, 如 U 的 d 领域范围内对象个数最大为 k 个, $k = (1 - f) * n$, 则 p 部分数据对象个数远大于 k。

这里距离函数的选取非常重要, 我们选取欧氏距离:

$$DIS(U, f) = \sqrt{\sum_{i=1}^{n} (o_i - f_i)^2}$$

定理 2: 如果 S 集中任意点 s 不是离群数据, $S \in T$, $s \in S$, f 与其第 $k+1$ 个最邻点距离远远大于 s 点与其第 $k+1$ 个最邻点的距离, f 是离群数据。

此定理非常重要, 它是本文采取的基于距离的离群数据挖掘的基础, 下面给出证明。

证明: 假使 f 不是离群数据, 则 $k+1$ 个最邻点与 f 点的距离 $DIS(k+1, f) \leqslant d$, 又 s 点与 $k+1$ 个最邻点距离 $DIS(k+1, s) \leqslant d$, 这与已知条件相矛盾, 因此 f 必定是离群数据。

3.2 算法描述

基于距离的离群数据的挖掘方法即第 k 个最近邻居法的基本思想是, 离群数据总是远离大部分的正常数据。基于这样的思想, 我们把数据中的每个记录看作是空间上的一个点, 计算每两点之间的距离, 找出与共它点相距距离最大的点, 该点所代表的数据即为离群数据。

第 k 个最近邻居方法的算法描述如下[9]:

```
Procedure FindUset(T, c) //根据离群置信度求离群数据
  Begin
  K= genvalue(e) //由离群置信度 c 求 k, k= int(c* m)
  For(i= 1; T[t_k, i] ≠φ; i++)
  Genvalue(t_i) //读取的 t_i 值
  For(j= i+ 1, T[t_k, j] ≠φ; j++)
  Genvalue(t_j) //读取的 tj 值
  DIS(t_i, t_j) //根据已定义的距离函数求 t_i, t_j 两点的距离
  For(l= 1; l< = K; l++)
  GenKset(DB(t_i, t_j), K) //将前 K 个距离数据保存
  Endfor
  Endfor
  For(l= 1; l< = K; l++)
  TopKset(DB(t_i, t_j)) //与上次保存的距离数据相比较
  Endfor
  Endfor
```

通过分析可以知道, 该算法运行过程中需要对数据库进行 $n*(n-1)/2$ 次遍历, 由于离群数据的数目远远小于正常数据, k 值远小于 n, 故此算法的时间复杂性近似地表示为 $0(n^2)$。

4 模拟实验

我们知道, 传统的数据挖掘方法在分析客户的信用等级方面做了许多有益的工作, 但不同的客户使用信用卡的模式有什么不同? 什么样的使用模式可能发生欺诈行为? 如何鉴别合法的用户及非法的用户? 这些问题的解决靠得是管理者的经验、能力和智慧以及事后的数据分析和处理工作, 很难及时、准确地预测相应的欺诈行为。而离群数据的挖掘重点是挖掘那些远离大部分正常数据的离群数据, 一般来说, 一个人在相当长的一段时间内, 其使用信用卡的模式应该是较为固定的, 一旦发现远离其习惯使用模式的行为存在, 系统应发出相应的警报, 以提醒、帮助管理者及时、准确地甄别欺诈[10~11]。下面通过一个信用卡恶意透支的例子说明基于距离的离群数据的挖掘方法在分析、识别和甄别各种欺诈行为方面的应用。

为了检测该算法的有效性, 用 Visual C++ 实现了此算法, 利用 http://www.amadon.ibm.com 所提供的模拟客户信用卡数据库进行模拟实验, 这些数据包括客户的地理分布信息、基本信息、信用卡使用信息等[12]。模拟实验结果表明, 使用该算法进行信用卡恶意透支风险的挖掘和预测工作是有效的, 通过 1986 年至 1995 年的十年中的实际欺诈数与用该算法得到的预测欺诈数的比较, 预测准确率为 88.5%, 模拟计算的结果如图 1 所示。

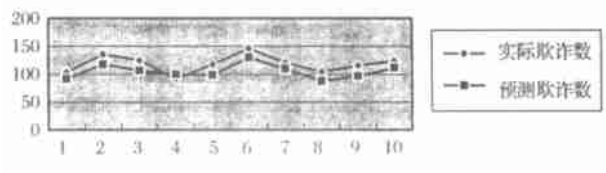

图 1 实际欺诈数与预测欺诈数的比较

从图 1 可以看出, 使用该方法能够帮助管理者了解恶意透支风险受那些因素的影响, 如何识别、分析和预测恶意透支风险, 以便在今后的信用卡销售、

服务和管理中有针对性地防止因信用卡恶意透支行为而给企业经营带来的损失。

5 结束语

本文提出了一种基于距离的欺诈风险分析方法。利用企业信息系统中数据库的数据, 在综合使用传统分析模型的基础上, 将数据挖掘技术应用于欺诈风险的识别、分析、评价中去。分析欺诈为什么会发生? 那些因素会导致欺诈? 欺诈风险主要来自于何处? 如何预测到可能发生的欺诈? 采取何种措施减少欺诈的发生? 通过评价欺诈风险的严重性、发生的可能性、欺诈风险指数及控制欺诈风险的成本, 汇总对各种欺诈风险的评价结果, 进而建立一套欺诈风险管理策略和监督体系, 设计并完善风险管理能力, 准确、及时地对各种欺诈风险进行监视、评价、预警和管理, 进而采取有效的规避和监督措施, 在欺诈风险发生之前对其进行预警和控制, 趋利避害, 从而使企业能够适应迅速变化的市场环境, 做好欺诈防范工作。

本文采用的距离是欧氏距离, 比较简单, 但精确程度欠佳, 可以考虑采用其他距离, 以便设计出更为精确、有效的算法, 降低算法时间复杂度和空间复杂度, 是今后需要进一步研究的问题[13~15]。此外, 用该方法进行欺诈风险的分析也存在一定的局限性, 它要求企业无任何超越市场规则的有失公允的交易行为、客户的信用卡数据必须是真实的和准确的、各种审计和检察机构的诚实等。对于因各种虚假的信用卡数据、信用卡客户与管理人员的串通合谋等欺诈行为而引发的欺诈行为的分析和预警是今后需要进一步研究的问题。

参考文献:

[1] 基特• 塞德格洛失[英]. 商务风险管理完全指南[M]. 沈阳: 沈阳出版社, 2001.

[2] 詹姆斯• 德阿克[美]. 企业的泛风险管理[M]. 长春: 吉林人民出版社, 2001.

[3] Marilyn Greenstein, Todd M Feinman. Electronic commerce: Security, Risk Management and Control[M]. McGraw-Hill Companies, 2000.

[4] 玛丽莲• 格林斯坦[美]. 电子商务的安全与风险管理[美]. 北京: 华夏出版社, 2001.

[5] Fayyad U M, Piatetsky-Shapiro G, Smyth P. Advance in knowledge discover and data mining[M]. California, The MIT Press, 1996, 1-25.

[6] 郑斌祥, 杜秀华, 席裕庚. 一种时序数据的离群数据挖掘新算法[J]. 控制与决策, 2002, (5): 324-327.

[7] 孙海洪, 夏克俭, 扬炳儒. 一种挖掘意外规则的快速算法[J]. 计算机工程与应用, 2001, (19): 49-51.

[8] 史东辉, 张春阳, 蔡庆生. 离群数据的挖掘方法研究[J]. 小型微型计算机系统, 2001, (10): 1234-1237

[9] Jiawei Han, Micheline Kamber. 数据挖掘概念与技术[M]. 北京: 机械工业出版社, 2001.

[10] 王越. KDD 方法在金融欺诈检测中的应用研究[J]. 计算机工程与设计, 2002, (5): 43-46.

[11] Ahn B S, Cho S SmKim C Y. The integrated methodology of rough set theory and artificial neural network for business failure prediction[J]. Expert System with Application, 2000, 18: 65-74.

[12] Huges. The complete database marketer: second generation strategies and techniques for tapping the power of your customer database[M]. Chicago, Irwin Professional, 1996.

[13] Alex Berson. 构建面向 CRM 的数据挖掘应用[M]. 北京: 人民邮电出版社, 2001.

[14] Charles X. Ling, Chenghui Li. Data Mining for Direct Marketing: Problems and solutions[C]. In Proceedings of KDD98: the AAAI-18 workshop on Knowledge Discovery in Databases, AAAI Technical Report, 1998.

[15] 刘同明. 数据挖掘技术及其应用[M]. 北京: 国防工业出版社, 2001.

A Research Method of Cheating Risks Based on Distance

LIU Bing-xiang[1], SHENG Zhao-han[2]

(1. School of Economics and Management, Southeast University, Nanjing 210096, China;
2. Institute of Management Science and Engineering, Nanjing University, Nanjing 210093, China)

Abstract: This paper tells the basic concept of the cheat and analyses the procreant reason of resulting in the cheat. The method of identification, analyzing and estimation of the cheat risks are discussed. It points out the existent problems of the tradition analyzing models in evaluating cheat risks. In the end, this paper puts forward to a research method of cheat risks based on distance and validates this method' s validity by a simulation experiment. The research work supplies a basis for further study of applying data mining for the enterprise cheating risks anglicizing.

Key words: cheating; distance; exceptional rules; outlier data; k-th nearest neighbor point