DOI:10.16208/j.issn1000-7024.2002.05.014

第23卷 第5期 计算机工程与设计 2002年5月
Vol.23 No.5 COMPUTER ENGINEERING AND DESIGN May. 2002

文章编号:1000-7024(2002)05-0043-03 中图分类号:TP311.13 文献标识码:A

KDD方法在金融欺诈检测中的应用研究

王 越[1,2], 曹长修[2]

(1. 重庆大学自动化学院, 重庆 400044; 2. 重庆工学院计算机科学与工程系, 重庆 400050)

摘 要: 在分析了金融事务中进行金融欺诈的现象后, 对传统的金融欺诈检测方法进行了分析, 并在此基础上, 提出了一种利用数据挖掘方法进行金融欺诈检测的模型, 并在此基础上利用该模型列举了方法运行的案例。

关键词: KDD; 数据挖掘; 金融欺诈; 决策树

Application research on detecting finance cheating using KDD method

WANG Yue[1,2], CAO Chang-xiu[2]

(1. Chongqing University Automation Institute, Chongqing 400044, China;
2. Chongqing Industrial Institute Computer Science and Engineering Dept., Chongqing 400050, China)

Abstract: After analyzing the phenomenon of finance cheating, this paper analyses the traditional method on detection of the finance cheating, and put forward to a model of data mining based on the above analysis to detect the finance cheating, finally, gives some examples based on the model.

Key words: KDD; data mining; finance cheating; decision tree

1 引 言

随着经济的发展,在金融领域的欺诈行为已经越来越多,为防止和检测金融欺诈所带来的费用也逐年增加。有调查表明,此项费用将继续呈上升的趋势。金融欺诈所带来的损失是巨大的,同时调查这些欺诈的费用也十分惊人。如何有效地减少这种损失,及时发现欺诈行为是非常重要的。

欺诈行为在金融服务领域非常普遍,比如保险业的金融欺诈就非常典型, 美国每年的保险欺诈高达100亿美元。典型的欺诈包括制造假的事故(如伤残欺诈)、财产索赔、制造虚假的医疗账单,等等。检测这些欺诈又相当困难, 需要非常熟练技巧的调查人员,通过查阅大量的相关记录,从一些蛛丝马迹中发现问题,然而与此同时,类似的欺诈行为仍在继续。

大型数据库管理系统是金融和保险系统使用的一种基本系统软件,在大型数据库系统中采用数据挖掘的方法是检测金融欺诈的一种先进的技术手段[1]。在大量的处理业务数据中对数据进行聚类分析找出相应的规则、规律、论断,再结合人的分析,是检测金融欺诈的一项有效的方法。本文第2部分对金融欺诈的现象进行了分析和分类,第3部分对传统的金融欺诈检测技术进行了分析,第4部分提出了采用KDD的数据挖掘技术的金融欺诈检测的完整方案,第5部分对该方法进行了总结。

2 金融欺诈的特征及类型分析

2.1 虚假的保险索赔

这种诈骗主要是通过制造假保险来骗取保金,以下是几个例子:

(1)某人将他的车投保,然后想法将车藏起来或将其卖掉,向保险声称车子丢失以骗取保险金。

(2)通过制造交通事故,比如将车子突然停下造成后面的车子追尾,然后要求医疗和财产赔偿,声称头部受伤,并由医生(同伙)出具医疗证明,骗取高额保险金。

2.2 健康保险诈骗

(1)进行本不必要的检查,如高额的心脏和肺部检

收稿日期: 2001-10-10

作者简介: 王越(1961-),男,北京人,副教授,重庆大学博士研究生,重庆工学院计算机与工程系主任,主要研究方向为数据库及其应用、KDD、计算机网络。 曹长修,教授,博士生导师,从事控制理论、计算机网络、人工神经网络等方面的教学和科研工作。

查，而实际上只不过是普通的感冒之类的疾病。

(2)医疗账单上写的是由资深的医生诊断而实际上是见习医生。

2.3 非保险类诈骗

(1)窃取信用卡或信用卡信息用来购买大宗商品。

(2)通过人为的手段使资产"升值"，然后使用这些升值的资产来进行诈骗活动。例如，几个人合伙买下一幢房子，然后转手将房子以高价卖给他们中的一个，这样房子的价值就升了上去，之后就可以以这幢房子进行借贷活动或转手卖给其他对此不怀疑的买主。

(3) 通过各种手段将非法所得的金钱存放到大量的账号中去，或者将这些黑钱转为合法收入。

通过以上分析可见欺诈行为有以下特点：

(1)表面合法性：任何欺诈行为在其表面上都是合法的，这就给检测欺诈带来了许多困难。

(2)有意性：欺诈行为都是预先设计的，而不是随机的。

(3)数据不合理性：因为从本质上讲欺诈是不合理的，所以任何欺诈在数据上或多或少地要显示出其不合理性，这就给采用数据挖掘技术进行欺诈的检测提供了基础。

3 传统的检测欺诈的技术

传统的检测金融欺诈的方法主要是依赖于计算机数据库应用系统支持的调查工作以及客户的受教育程度。传统的检测欺诈技术的工作流程如图1所示。

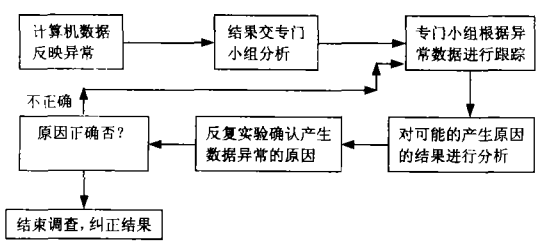

图1 传统的检测欺诈技术工作流程图

从传统的欺诈检测方法来看有以下缺点：

(1)滞后性；

(2)不准确性；

(3)不及时性。

4 数据挖掘检测欺诈技术

数据挖掘是用来处理大型数据库的，因此它提供了对金融欺诈进行检测分析的环境。决策树作为知识发现的算法首先在ACSys数据挖掘系统(Williams和Huang 1996)中使用[2]。决策树是利用信息论中的互信息(信息增益)寻找数据库中具有最大信息量的字段，建立决策树的一个结点，再根据字段的不同取值建立树的分枝；在每个分枝子集中重复建立树的下层结点和分枝的过程，直到生成一个完整的决策树。决策树的实现需要包含3种操作：树的生长、树的评估、树的应用。树的生长阶段通常在一并行体系结构中实现分类－克服策略，在一组训练例的基础上建立一个完整的决策树；树的评估阶段利用测试例集合来评价生成的树，在这一阶段，需要对树做出适当的修剪，并选取不同的测试例集合来对树的性能进行测试并进行修剪；树的应用是将最后生成的树应用于未知的数据。

从数据挖掘的观点来看，对于金融欺诈将从以下几个方面进行分析。

(1)异常数据：相对于自身的异常数据，相对于其他群体的异常数据(检测比较困难)。

(2)无法解释的关系：如在医疗账单中，相当多的人有相同的医生或同一地址。

(3)通常意义下的欺诈行为：一旦一个欺诈行为被证实，那就可以使用它来帮助确定其他可能的欺诈行为。这些事务可能已经发生过并且被处理过了，或在将来要被处理，或在将来可能发生，或兼而有之。这种类型的分析叫做“预测数据挖掘”。应用这种技术需要通常要用到3个数据挖掘工具：回归、决策树、神经元网络。

有用的预测可以被合并，加入到历史数据库中并用来帮助寻找相近未被发现的案例。随着成功案例的积累，预测系统的质量和可信赖度会大大增强。

这种方法相对于以往的警示系统而言的优点，是它的可信赖度可以被统计评估和证实。如果可信任度很高，那么大多数的调查可以集中处理实际的欺诈事件，而不是在大量似是而非的案例上寻找。

(4)数据挖掘的目标与响应的技术之间的关系见表1。

表1 数据挖掘检测技术

任务	目标	数据挖掘技术
发现异常数据	检测全局异常纪录 检测多发生事件的值 检测纪录之间的连接关系	异常分析
通常的欺诈行为特征	基于历史数据找到标准，如检测欺诈行为的规则	预测模型
	纪录下可能或类似欺诈事务	
证实无法解释的关系	检测具有不正常值的纪录	聚类分析和异常分析
	确定嫌疑人的图表	聚类分析
	检测相同或相近的纪录	
	检测纪录之间的非直接的联系	社会关系网络和连接分析
	检测混合的异常纪录	联合分析，序列分析

5 实例分析

以下系统是根据某银行信用卡部的交易记录，利用 INFORMIX FOR UNIX 数据库、POWERBUILDER 6.5 及 VISUAL C++ 6.0 语言编制的一个小型数据挖掘系统，并给出了数据挖掘对金融欺诈检测的应用实例。

这里给出一个实例，说明使用数据挖掘技术如何来检测信用卡诈骗的步骤。

(1) 首先要建立一个包含事务记录的数据集。这里的数据大约包含4,736个信用卡交易的记录，诈骗比率可以设为1/20。

(2)分析:图2显示了一个标准的数据挖掘的模型[3]:

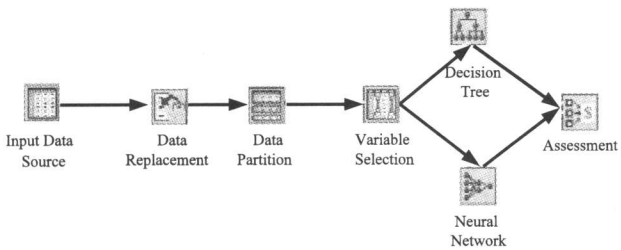

图 2 数据挖掘模型

Input Data Source 节点用来指定数据源；Data Replacement 节点用来处理缺欠的数据；Data Partition 节点用来生成训练和有效的数据集；Variable Section 节点用来决定使用哪些变量和放弃那些无用的变量；这里还用到了两个模型：决策树(Decision Tree)和神经元网络(Neural Network)，这两个模型处理的结果送交打分(Assessment)节点来比较，打分节点比较的结果可以用图表的形式显示出来，图3是打分节点关于神经元网络的 LiftChart 图。

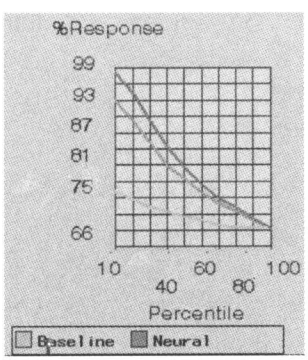

图 3 神经元网络模型

最上面的曲线是神经元网络模型，中间的曲线是决策树，最下面的曲线代表基线。从曲线上可以看出，在10%的数据中，神经元网络模型的诈骗预测数目将近诈骗总数的98%，决策树模型的诈骗预测数目将近诈骗总数的93%。在40%时，神经元网络和决策树分别达到80%和78%。从中可以看出这里的神经元网络模型要优于决策树模型。

决策树的优点是比较简单，并且可以比较直观地显示出来。图4是一个决策树的树型显示。

(3)结论:使用数据挖掘技术允许对交易进行欺诈

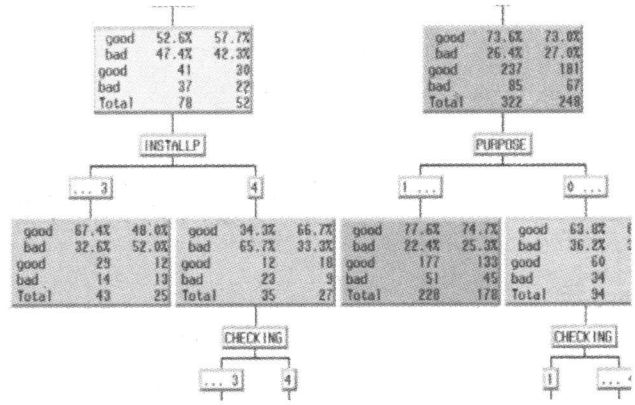

图 4 决策树的树型

预测，利用一些模型，如决策树可以方便地将结果以可视化的形式表达出来。神经元网络的结果虽然不是可以很清晰地看出来，但它的结果往往要优于决策树的结果。打分节点可以对使用的模型加以评比，帮助选择最优的解决模型。每个模型建立起来以后，都可以将生成的代码使用到新的数据中去预测欺诈行为，调查工作可以在预测的范围内展开，这可以节省大量的时间和费用。

6 结束语

金融欺诈的范围很广，充斥于很多的金融业务当中，有时很难被传统的技术和方法检测和发现，因此对金融公司是个很大的威胁。在大量的实际工作中，传统的技术非常费时，并且不能覆盖众多的欺诈类型，所以很多的公司在寻找IT技术的帮助，以避免给其单位带来损失。

利用数据挖掘技术来检测欺诈，可以提供完整的解决方案。从我们的实践来看，利用数据挖掘技术不但可以节约人力并且可以节省时间，这样可以迅速地使犯罪分子绳之于法。就目前来看，如何提高挖掘准确性和挖掘速度，是本课题应深入研究及实践的问题。对此方法的应用研究在金融保险领域有着深远的现实意义。

参 考 文 献:

[1] 王能斌.数据库系统原理[M].北京：电子工业出版社，1999.

[2] Fayyad M, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery: an overview [M]. Advances in Knowledge Discovery and Data Mining. Menlo Park, CA: AAAI Press, 1996.1-36.

[3] Michalski R, Stepp R. Automated construction of classification: conceptual clustering versus numerical taxonomy [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1983,5 (4): 396-409.