聚类与关联规则在信息舞弊识别中的应用①

幸莉仙, 黄慧连

(华北电力大学大学 经济管理系, 保定 071003)

摘 要: 针对现代电子数据迅速膨胀, 传统的审计方式已经无法应对海量的业务数据, 试图将数据挖掘中的聚类和关联规则算法引入审计领域. 在研究聚类与关联规则算法的含义及相关算法—K-Means 和 Apriori 算法的基础上, 提出了一种基于聚类与关联规则的审计模型, 并以某市城镇医疗保险的审计为例, 首先利用聚类分析进行数据筛选, 然后利用关联规则挖掘海量数据之间潜在的关系, 为审计提供线索. 文章通过案例分析为数据挖掘在信息舞弊识别领域的应用提供参考.

关键词: 信息舞弊; 关联规则; Apriori 算法; 聚类; K-Means; 数据挖掘; 审计

The Application of Clustering and Associate Rule Mining to Fraud Information Identification

XING Li-Xian, HUANG Hui-Lian

(School of Business and Administration, North China Electric Power University, Baoding 071003, China)

Abstract: Considering that with the rapid expansion of electronic data, the traditional audit approachs can not cope with vast business data, this paper intend to introduce the Clustering and Association Rule Mining in the audit fields. Based on the study of the meaning of Clustering and Association Rule Mining and their Algorithm—K-Means and Apriori, this article proposed an audit model which is based on the Clustering and Association Rule Mining, at the same time, taking the audit of medical insurance of some a city as an example, it detailed first how to use the Clustering to filter data, then how to mining the potential relationships in vast data so as to determine the audit priorities and audit clues.Through the case, the article is committed to provide a reference for the application of data mining in the fraud information identification.

Key words: fraud indormation; association rule mining; apriori; clustering; K-means; data mining; audit

随着计算机技术、网络技术、通信技术的飞速发展和普及, 财务管理系统, 企业资源计划(ERP)、供应链管理系统等一系列计算机信息系统开始广泛应用于各行各业以及各企事业单位. 这些技术和产品引领社会进入了一个 IT 管理的环境, IT 技术为企业带来巨大经济效益的同时也给审计人员带来了巨大的挑战. 首先, 各种计算机软件的应用使得各行各业积累了相当规模的、不同领域、不同种类、不同存储形式的海量数据,使得传统审计技术和方法无法应对; 其次, IT 技术也给不法分子带来了更加先进的信息舞弊和造假手段.

目前信息舞弊呈现出智能化、复杂化、主体多元化、隐蔽化、实时化等一系列新的特点[2,4]. 将数据挖掘技术引入审计领域已经成为当前审计发展的一个重要趋势. 本文将数据挖掘算法中的聚类算法与关联规则算法应用于会计舞弊信息识别的研究, 首先通过聚类分析来获取可能存在舞弊现象的数据集, 然后试图利用关联规则挖掘海量数据之间的内在联系, 发现会计舞弊线索, 从而锁定审计重点, 为审计提供依据.

1 聚类分析

聚类(Clustering)是数据挖掘领域最为常见的方法之一, 它最核心的思想就是“物以类聚”, 将具有相似性

① 收稿时间:2012-05-07;收到修改稿时间:2012-06-26

的对象分为一类, 不相似的对象分为不同的类. 聚类也是一种分类方法, 但是它是一种典型的无监督学习过程, 它与其他分类算法最典型的区别在于其他分类算法首先需要实现知道所依据的数据特征, 而聚类是要找到这些数据特征, 因此在很多情况下聚类是一种有效的数据预处理和筛选的方法[7]. 本文利用聚类算法的这一特性来排除那些舞弊可能性较小的数据, 从而减少数据量, 提高后续关联规则算法的效率, 也剔除大量的弱关联规则.

目前聚类分析的方法有很多, 包括基于层次的聚类、基于划分的聚类、基于密度的聚类以及基于网络的聚类等. 比较经典和常用的聚类方法是基于划分的 K-Means 方法. K-Means 算法以 K 为参数, 把所有对象分成 K 类, 使得同一类别中的样本具有较高的相似度, 而不同类别的相似度较低[8]. K-Means 算法的具体流程描述如下:

输入: n 个数据对象集合 $x_1, x_2, \cdots, x_n$, 分类数 K.

输出: K 个聚类中心以及 K 个聚类数据对象集合 S_j.

(1) 从所有对象中随机选择 K 个样本点 $x_1, x_2, \cdots, x_k$ 作为初始聚类中心 $C_1, C_2, \cdots, C_k$.

(2) 计算各个样本点到各个聚类中心的距离, 并将其分配到距离最小的类簇中.

(3) 计算每个类簇中所有对象均值, 形成新的聚类中心.

(4) 重复第(2)和第(3)步, 直到数据的划分不再变化或者达到最大的迭代次数.

在识别会计信息舞弊特征时, 通过聚类分析, 将大量数据进行分类, 对于那些不存在舞弊特征的数据可能聚成一类, 通过排除这些数据就可以进一步锁定审计重点, 缩小审计范围, 提高关联规则挖掘的效率和精度.

2 关联规则算法

1993 年 Agrawal 等人首先提出了交易数据库中不同商品之间的关联规则挖掘, 并逐渐引起了专家、学者的重视, 在目录审计、销售追加、仓储规划等多个领域得到广泛应用. 关联规则挖掘是数据挖掘的一项重要研究内容, 其目的就是在海量的数据中发现数据项之间的潜在关系, 反映数据项之间的密切程度[1,3]. Agrawal 等人设计关联规则的经典算法——Apriori 算法, 它是层次算法的基础, 是最典型的层次算法, 其核心技术为其它各类关联规则挖掘算法广泛采用, 是一种最有影响的挖掘布尔关联规则频繁集项的算法. Apriori 算法分为挖掘频繁项目集和产生关联规则两个阶段.

2.1 挖掘频繁项目集

Apriori 算法采用逐层搜索的迭代方法, 扫描数据库产生频繁项集. 首次扫描数据库生成 1 项频繁项集记为 L1, 然后然后根据 L1 再次扫描数据库产生 2 项频繁项集 L2, 然后根据 L2 扫描数据库生成 3 项频繁项集, 一次循环下去, 直到找不到频繁项. 主要伪代码如下:

```
//输入事务数据库 D 和最小支持度 min_support
//...数据准备阶段
L1=generate_1_frequentitem_set();//第一次扫描数据库, 生成 1 项频繁项
For(i=1;Li-1!=Φ;i++) do
{
    Ci=generate_i_Candidate_frequentitem_set(Li-1);//根据 i-1 项频繁项集 Li-1 生成 i 项候选频繁项集 Ci
    For each t ∈ D//扫描数据库, 统计 Ci 中个候选频繁项的支持度
    {
        For each c∈Ci //对于 Ci 的任意候选频繁项 c, 如果是事务 t 的子集, 则 c 的支持度加一
        If(c=Subsetof(t))
        {
            c.support++;
        }
    }
    Li={c∈Ci|c.support>=min_support}
}
Return L=U_i L_i
```

事实上, 频繁项集具有这样的性质: 在给定的事务数据库中, 任意频繁项集的子集都是且必须是频繁项集. 利用这一性质算法在判断 Ci 中的候选频繁项集是否为频繁项集时可以利用此性质进行修剪.

2.2 产生关联规则

挖掘出项目频繁项以后要获得的关联规则就比较简单了. 利用上述计算可信度的公式, 对于大于最小可信度的规则即我们挖掘出来的强关联规则, 产生关联规则的伪代码如下:

```
For each L in Lk(K>=2)//对于所有 K 项频繁项 L
{
    H={L 中规则的结论，该算法中规则的结论只有一个项};
    For each h= {h∈subset(L)|H 不属于 h}//对于任意属于 L 且不含 H 的项目子集 h，计算其可信度
    {
        Conf_h=support(h)/sopport(H);
        If(Conf_h>=min_confifent)//如果 L 的项目子集 h 的可信度大于最小可信度，则 h=>H 为一条强关联规则，输出规则及其可信度
        Output h=>H,confidengce=Conf_h;
    }
}
```

3 应用研究

3.1 事件背景

在对某省 X 市城镇职工医疗保险专项审计调查中，审计组发现当地城镇职工医疗保险业务数据高度信息化，后台数据库为 Oracle10g，全库备份数据库容量达 13G，其中仅处方明细表记录条数多达 2100 万多条，面对如此海量数据，审计组不可能对每一组数据进行一一核对，只能从这些数据中抽取重点详细调查，但仅仅依靠审计人员的经验很难从浩瀚的业务数据中快速找出审计重点所在，将关联规则对整体数据进行分析，挖掘数据之间的潜在规则，确定审计重点.

3.2 业务描述

审计组通过学习当地医保相关文件发现，参保人员中离休及革命伤残军人实行实报实销制，全额财政拨款，这部分人群中可能存在定点医疗机构串换药品、参保人员医保卡出租出借，一人参保多人享受等违规行为. 为了解该市参保人员是否存在这种现象，审计人员需要对以往的医疗数据进行核实.

3.3 数据准备

审计组采集该市医保数据管理系统的底层数据，其中包含与参保人员有关的所有信息，在前期整理、分析中，确定医保审计的两张中间表：医保人员信息表、就诊明细表、处方明细表、医疗费用结算明细表. 各表主要字段信息分别见表 1、表 2、表 3:

表 1 医保人员信息表

身份证号	姓名	性别	出生日期	家庭住址	报销比例	联系方式
Varchar	Varchar	Varchar	Varchar	Varchar	Number	Varchar

表 2 就诊明细表

身份证号	定点医疗机构名称	地点	药品金额	药品名称	药品数量	就诊科室	就诊时间
Varchar	Varchar	Varchar	Number	Varchar	Number	Varchar	Varchar

表 3 医疗费用结算明细表

身份证号	医疗费用结算金额	医疗类别	结算时间	结算定点医疗机构
Varchar	Number	Varchar	Varchar	Varchar

为将三张表的信息汇总，又新建一张“审计分析表”，表结构如下表 4 所示:

表 4 审计分析表

身份证号	姓名	医疗费用总金额	医疗费结算金额	药品总金额	药品总数	就诊次数	报销比例
Varchar	Varchar	Number	Number	Number	Number	Number	Number

3.4 聚类分析

由于数据量过于庞大，而在审计中舞弊主体的数量相对来说会比较少，这样如果不对数据做任何处理舞弊特征可能很难挖掘出来. 采用 K-Means 算法根据医疗费用总金额、医疗费用结算金额、药品总金额、药品总数量、就诊次数这五个属性对数据进行聚类分

析, 通过尝试将数据分成了 8 类, 保留其中数值较大的四个分类进行下一步的关联规则挖掘.

3.5 关联规则挖掘

采用 Java 编程按照上述思路设计 Apriori 算法代码, 将 min_support=0.1 和 min_confidence=0.6;选择医疗费用总金额、医疗费用结算金额、药品总金额、药品总数、结算次数、就诊次数和报销比例属性作为选择的数据项, 对筛选出来的数据进行离散化整理后进行关联规则挖掘, 经过整理发现如下可疑规则如下表 5 所示:

表 5 关联规则挖掘结果

编号	关联规则	可信度
1	药品总金额>1 万，药品数量>500 -----> 报销比例>=85%	0.866
2	医疗费用总金额>2 万------>报销比例>=85%	0.832
3	医疗费用结算金额>1 万，就诊次数>20 ------> 报销比例>=85%	0.823

3.6 挖掘结果分析

对表 5 中的三条关联规则进行分析, 我们发现报销比例大于 85%的参保人员在医疗费用、药品的费用、药品数量、就诊次数都处于较高的水平, 因此审计初步判断该地区医疗保险参保人员可能存在串换药品、租借医保卡、一人参保多人享受等违规现象. 审计人员可以将这类参保人员作为重点审计对象, 透过这些线索提示进行分析取证, 最后落实这种现象的真伪.

3 结语

将数据挖掘技术与海量数据下的审计相结合将成为未来审计的发展方向之一. 本文介绍了数据挖掘中的聚类和关联规则算法, 并重点讨论了 K-Means、Apriori 算法, 将 K-Means 与 Apriori 算法应用于医保审计, 发现医保审计数据之间某些隐含的规则, 从而确定审计重点, 提高审计效率. 此外, 对于那些有价值的新规则可以添加到审计规则库中, 指导审计; 随着审计规则库的日益扩大, 使得审计经验得到丰富, 提高审计质量.

但是将关联规则应用于信息审计还存在如下几个方面的问题:

① 就关联规则算法来说, 目前关联规则算法的执行效率不高. 每生成一个频繁项算法都需要扫描数据库, 这种成本是非常高的, 因此算法的改进是我们今后研究的方向之一.

② 关联规则算法首先要设置支持度和可信度, 这就要求某种行为成为一种达到一定数量或者比例的群体性行为时, 才能作为一种规则被挖掘出来. 在案例中, 只有当达到一定数量参保人员有串换药品、参保人员医保卡出租出借, 一人参保多人享受等违规行为时, 表 5 中的规则才能被挖掘出来. 也就是说, 当参保人员中只有少数人有这种舞弊行为时, 表 5 中的规则是不可能被挖掘出来的. 此时, 数据挖掘中的孤立点算法成为挖掘这种现象的有力方法.

因此, 一方面我们要完善挖掘算法使之适应审计的需要, 另一方面, 我们也应该将各种数据挖掘算法相结合, 从而更好、更全面地为审计提供辅助.

参考文献

1 廖芹,郝志峰,陈志宏.数据挖掘与数学建模.北京:国防工业出版社,2010.292–300.
2 刘汝焯,等.审计线索的特征发现.北京:清华大学出版社,2009.47–67.
3 元昌安,邓松,李文敬,刘海涛,等.数据挖掘原理与 SPSS Clementine 应用宝典.北京:电子工业出版社,2009.176–189.
4 韩学鸿,贾瑞敏.数据挖掘技术的应用研究综述与启示—在会计舞弊识别研究中的应用.生产建设,2009:119–120.
5 李胜.基于关联规则的审计特征智能提取的应用研究.北京:北京交通大学,2006.
6 陶振海,谢凯年.审计数据多维关联规则挖掘算法.计算机应用与软件,2008,9(25):161–168.
7 Han JW, Kamber M.数据挖掘概念与技术.北京:机械工业出版社.第 2 版.2007.
8 张玉芳,毛嘉莉,熊忠阳.一种改进的 Kmeans 算法.计算机应用,2003,23(8):31–34.