Apriori算法在税务系统中的应用

王 敏[1],刘希玉[2]
(1.山东师范大学 信息科学与工程学院,山东 济南 250014
2.山东师范大学 管理与经济学院,山东 济南 250014)

摘 要:为提高对税务稽查历史数据的管理水平,为税务稽查提供有力的决策支持,分析了数据挖掘技术的基础以及税务稽查工作中引入数据挖掘技术的必要性,给出了关联规则挖掘的基本概念,提出了一种改进的Apriori算法。使用WEKA工具,实现了关联规则挖掘的可视化和生成结果按"支持度-可信度"形式的可视化,为基于频繁集的交互式挖掘提供了方便、友好的界面。对历史稽查数据中纳税人采用的主要违法违章手段之间的关联关系进行了数据挖掘,得到了一些合理的规则,对稽查工作有一定的指导意义。

关键词:数据挖掘;Apriori算法;WEKA;税务稽查;关联规则
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2009)11-0175-04

Application of Apriori Algorithm in Tax System

WANG Min[1],LIU Xi-yu[2]
(1.Department of Information Science and Engineering,Shandong Normal University,Jinan 250014,China;
2.Department of Management and Economics,Shandong Normal University,Jinan 250014,China)

Abstract:To increase the level of inspection historical data management and provide a strong tax reform policy support, it analyzes the basic of data mining technology and the necessity of data mining technology in tax inspector's work and gives the basic concepts of mining association rules. It uses an improved Apriori algorithm and WEKA visualizes generated results in support-confidence form, and provides an easily accessible and user-friendly interface for interactive mining based on frequent itemsets. Uses it into mining the association rules of tax offence of the historical data and receives some reasonable rules. These rules have guiding significance in inspector job.

Key words:data mining; Apriori algorithm;WEKA;tax reform;association rule

0 引 言

数据挖掘是从大量的数据中,获取潜在有用信息和知识的一种技术[1]。数据挖掘融合了人工智能、数据库、机器学习等各个学科的理论和技术。它通过各种方法从数据库中提取细节数据,然后进行综合、加工和集成,以适当的数据结构(如数据仓库)进行存储,然后从中提取隐含的、先前未知的、对决策有潜在价值的知识或规则。

随着税务信息化建设的发展,各级税收数据库的规模和数据量在迅速膨胀,各个系统中存储了大量的涉税信息。如何将这些历史静态的数据变成动态且具有分析决策价值的信息已成为当前急需解决的问题。文中针对税务系统大量的原始数据,尝试把数据挖掘中的相关方法和技术,应用到税务稽查中,通过纳税户的各种基本信息以及在纳税过程中的各种行为和数据,利用关联规则数据挖掘来发现最可能存在纳税问题的纳税户,为税务稽查的决策提供强有力的支撑,具有一定的实用价值和应用前景。

收稿日期:2009-03-03;**修回日期**:2009-06-12
基金项目:国家自然科学基金重大项目(60873058,60743010);山东省自然科学基金重大项目(Z2007G03)
作者简介:王 敏(1985-),女,硕士研究生,研究方向为数据挖掘与智能优化算法;刘希玉,"泰山学者",教授,博士生导师,研究方向为数据挖掘与人工智能。

1 税务稽查工作中引入数据挖掘技术的必要性

1.1 税务稽查工作现状

近年来,税务稽查工作通过发挥监督、惩处、教育和收入等职能,有效地维护了税法的严肃性和税收经济的正常秩序。但离新形势下税务稽查工作的要求还有一定的距离,特别是稽查管理的精细化、科学化程度还不高,主要表现在以下几个方面:

(1)税务稽查选案缺乏科学性、准确性。当前涉税

违法犯罪日益多样化、隐蔽化,传统的选案手段已经制约了稽查案件的有效性、针对性。

(2)稽查实施重点不突出。由于稽查选案的盲目性,制定稽查计划时,只能每年查一定的比例,几年轮一遍,这样就使检查的户数过多,导致日常稽查难以查深查透。

(3)数据分析手段单一,应用软件间缺乏信息共享。目前在稽查局运行的举报管理系统、中国税收征管信息系统等都是相互独立的系统,系统间缺少信息共享。并且这些系统对数据的分析也仅停留于简单的统计和报表功能,缺乏对大规模征管及涉税数据的加工分析[2]。

(4)电子商务对税务稽查的挑战。网络经济不仅提供了集信息流、物流和资金流于一体的商务交易模式,也对以会计信息系统为客体的税务稽查产生了极大冲击。具体表现为:税务稽查环境、范围、线索、技术方法以及税务稽查主体素质的改变等[3]。如何利用先进的计算机技术对电子商务进行"网上稽查",成为各级稽查部门需要思考和探索的问题。

1.2 信息资源整合为税务稽查提供数据分析平台

随着中国税收征管信息系统(CTAIS)的全面推行,以及整合主体应用系统的逐步实施,税务系统的信息化正由扩张走向信息资源整合和数据集中管理的阶段[4]。数据整合使业务数据集中处理迈向新台阶,不仅带来了数据处理平台的全面提升以及系统间信息交换共享的实现,而且也为税务稽查工作的数据分析带来了新的契机,给稽查工作提供了一个统一的、全面涵盖纳税人涉税信息的数据平台,使前沿的数据分析、挖掘技术有了发挥作用的基础,开辟了数据信息向知识信息转变的通道。

2 关联规则及相关概念

关联分析是为了发现大量数据中项集之间有趣的关联或相关联系[5]。设 $I=\{i_1,i_2,\cdots,i_m\}$ 是项的集合。任务相关的数据 D 是数据库事务的集合,其中每个事务 T 是项的集合,使得 $T\subseteq I$。每一个事务有个标识符 TID。关联规则(associations rule,AR)定义为 $A\rightarrow B$,其中 A 包含于 I,B 包含于 I,且 $A\cap B=\varnothing$。关联规则具有如下两个重要的属性:

(1)支持度:$\text{support}(A\rightarrow B)=P(A\cap B)$,即 A 和 B 这两个项集在事务数据库 D 中同时出现的概率。

(2)置信度:$\text{confidence}(A\rightarrow B)=P(B\mid A)$,即项集 A 出现的事务数据库 D 中,项集 B 也同时出现的概率。

同时满足最小支持度阈值(min_sup)和最小置信度阈值(min_conf)的规则叫做强规则。

项的集合叫项集(itemset),包含 k 个项的项集叫 k-项集。项的出现频率是包含项集的事务数,简称为项集的频率、支持计数或计数[6]。如果项集的出现频率大于或等于 min_sup 与 D 中事务总数的乘积,则项集满足最小支持度 min_sup。如果项集满足最小支持度,则称它为频繁项集(frequent itemset),频繁 K 项集的集合通常记为 L_k。

在事务数据库中挖掘项之间的关联规则,就是找出支持度和置信度分别大于用户给定的最小支持度(min_sup)和最小置信度(min_conf)的关联规则[7]。

3 税务稽查数据的关联规则挖掘

3.1 关联规则挖掘算法

首先给出 Apriori 算法,Apriori 算法描述如下[8]:

输入:事务数据库 D;最小支持度阈值 min_sup。

输出:D 中的频繁项集 L。

方法:

1) $L_1=\text{find\_frequent\_1-itemsets}(D)$;
2) for $(k=2;L_{K-1}\neq\varnothing;k++)\{$
3) $C_k=\text{apriori\_gen}(L_{K-1},\text{min\_sup})$;
4) for each transaction $t\in D\{$
5) $C_t=\text{subset}(C_k,t)$;
6) for each candidate $c\in C_t$
7) c.count++;$\}$
8) $L_k=\{c\in C_k\mid c.\text{count}\geqslant\text{min\_sup}\}\}$
9) return $L=\bigcup_k U_k$;

procedure apriori_gen((L_{K-1}: frequent(k-1)-itemsets; min_sup: minimum support threshold)

1) for each itemset $l_1\in L_{k-1}$
2) for each itemset $l_2\in L_{k-1}$
3) if$(l_1[1]=l_2[1])\wedge(l_1[2]=l_2[2])\wedge\cdots\wedge(l_1[k-2]=l_2[k-2])\wedge(l_1[k-1]=l_2[k-1])$then$\{$
4) $c=l_1\cup l_2$
5) if has_infrequent_subset(c,L_{k-1})then
6) delete c;
7) else add c to C_k
8) $\}$
9) return C_k

procedure has_infrequent_subset(c: candidate k-itemset; L_{K-1}: frequent(k-1)-itemset)

1) for each (k-1)-subset s of c
2) if s不属于 L_{K-1}
3) return true;
4) return false;

上面的算法中 C_k(L_{k-1} 直接生成的)到 L_k 经过了两步处理,第一步根据 L_{k-1} 进行裁剪,第二步是根据 minimum 进行裁剪[9]。当经过连接生成 K 维数据项集时,判断它的 $K-1$ 维子集是否存在与 L_{k-1} 中,如果不存在就直接删除。这样每生成一个 K 维数据项集时就搜索一遍 L_{k-1}。改进算法的思想就是只搜索一遍 L_{k-1} 就可以了。当所有连接完成的时候,扫描一遍 L_{k-1},对于 L_{k-1} 任意元素 A,判断 A 是否为 C_k 中元素 c 的子集,如果是,对子集 c 进行计数。也就是统计 L_{k-1} 中包含 C_k 中任意元素 c 的 $k-1$ 维子集的个数。最后根据 c 进行裁剪,如果 $c < K$,则删除。

算法的主体不变,apriori_gen 函数改进如下,has_infrequent_subset 函数不再需要。

procedure apriori_ gen((L_{K-1}: frequent(k - 1) - itemsets;min_ sup:minimum support threshold)

1) for each itemset $l_1 \in L_{k-1}$

2) for each itemset $l_2 \in L_{k-1}$

3) if($l_1[1]=l_2[1]) \wedge (l_1[2]=l_2[2]) \wedge \ldots \wedge (l_1[k-2]=l_2[k-2]) \wedge (l_1[k-1]=l_2[k-1])$then{

4) $c=l_1 \cup l_2$

5) for each itemset $l_1 \in L_{k-1}$

6)for each candidate $c \in C_k$

7) if l_1 is the subset of C_k then

8) c. number + + ;

9) $C_k=\{c \in C_k \mid c.number=k\}$;

10) return C_k

一旦数据库 D 中的事务找出频繁项集,则关联规则可以产生如下:

(1)对于每个频繁项集 l,产生 l 的所有非空子集。

(2)对于 l 的每一个非空子集 s,如果 support_count(l)/support_ count(s) $\geqslant$ min_ onf,则输出规则"$s \Rightarrow (l-s)$"。其中,min_ conf 是最小置信度阈值。

产生关联规则 $A \rightarrow B$ 可作为挖掘税务稽查数据的依据,为税务稽查工作提供决策支持。

3.2 举　例

假设 L_{k-1} = {{1,2,3},{1,2,4},{2,3,4},{2,3,5},{1,3,4}},求 L_k。

由 L_{k-1} 得到 C_k = {{1,2,3,4},{2,3,4,5}}。

原算法:首先得到{1,2,3,4} 的子集{1,2,3},{1,2,4},{2,3,4},{1,3,4}。然后判断这些子集是不是 L_{k-1} 的子集,如果都是,则保留,否则删除。这里保留{1,2,3,4},{2,3,4,5} 则应该删除。得 C_k = {{1,2,3,4}}。

改进算法:首先从 L_{k-1} 中取出元素{1,2,3},扫描 C_k 中的元素,看{1,2,3} 是不是 C_k 中元素的子集,{1,2,3}是{1,2,3,4}的子集,{1,2,3,4}的计数加 1,{1,2,3}不是{2,3,4,5}的子集,计数不变,经过对{1,2,3}处理后得到计数{1,0};然后看{1,2,4},{1,2,4}是{1,2,3,4}的子集,而不是{2,3,4,5}的子集,计数变为{2,0},考察{2,3,4},{2,3,4}是{1,2,3,4}的子集,也是{2,3,4,5}的子集,计数变为{3,1};{2,3,5}不是{1,2,3,4}的子集,是{2,3,45}的子集,计数变为{3,2};{1,3,4}是{1,2,3,4}的子集,不是{2,3,4,5}的子集,计数变为{4,2},数据扫描完毕。此时 $K=4$,只有第一个元素的计数为 4,为高频数据项集得到 C'_k = {{1,2,3,4}}。

3.3 算法分析比较

下面对原算法和改进算法的性能进行比较。

L_{k-1} 中的数据项集的个数记为 $|L_{k-1}|$,C_k 中数据项集的个数记为 $|C_k|$,C_k 中元素的自己个数设为 n_i,其中 $i=1 \sim |C_k|$。这里只分析从 C_k 到 C'_k 的处理,原 Apriori 算法从 C_k 中取元素,然后求该元素的子集,判断该子集是否在 $|C_k|$ 中,需要进行的运算为 $|C_k| n'_i(|L'_{k-1}|)$ 次,$1 \leqslant |L'| \leqslant |L'_{k-1}| \leqslant |L_{k-1}|$,$1 \leqslant n_i' \leqslant n_i$。而改进算法是从 L_{k-1} 中选取元素,看是不是 C_k 中元素的子集,对 C_k 中数据项集子集个数进行统计[10]。需要进行的计算是($|L_{k-1}|+1$) * $|C_k|$ 次。如果 $n'_i=1$,就是每次只取 C_k 中数据项集的一个子集就可以判断该数据项集,则两个算法的效率基本相同,但是这种情况很少出现,从而大部分情况下,改进的算法效率要高于原算法。

3.4 关联规则挖掘过程

应用上述改进的算法对税务稽查数据进行频繁模式和关联规则挖掘。稽查数据库中的稽查数据如表 1 所示。其中 TID 为违章案件编号,A、B、C、D 等为纳税人违法违章类型代码。

表 1　数据库的违法违章数据记录

TID	违法违章类型
1	B、E、A、B
2	A、B、C、E
3	A、B、C、D、F
4	E、A、F、D
5	A、B、C、G
6	F、D、C、C
7	B、E、F
8	B、C、F、H
9	H、C
……	……

挖掘频繁项集的过程如下:首先对稽查数据进行数据预处理,即将稽查数据进行排序并删除重复项,将

数据整合成能被挖掘算法利用的数据，并存入稽查数据库 D 中[11]，表 1 的数据经预处理后转化为表 2。

表 2 预处理后的数据挖掘库中的数据

TID	违法违章类型
1	A、B、E
2	A、B、C、E
3	A、B、C、D、F
4	A、D、E、F
5	A、B、C、G
6	C、D、F
7	B、E、F
8	B、C、F、H
9	C、H
…	…

数据整理完毕，使用 WEKA[12]进行实验。在具体的挖掘过程中，根据给定的最小支持度 α(min_supp) 和最小可信度 β(min_conf) 的不同得出如下结果。例如在 $\alpha=0.17$，$\beta=60$ 时，得到 62 条关联规则(见表 3)。

表 3 关联规则挖掘结果

关联规则	支持度	可信度 %
AG→K	0.25	78.74
GT→K	0.24	81.26
GK→T	0.24	62.8
MX→R	0.23	72.63
MR→X	0.23	62.8
JW→N	0.22	81.6
NW→J	0.21	80.3
AX→C	0.21	63.34
BY→D	0.2	83.74
XZ→M	0.2	82.43
…	…	…

随着税务信息化建设的进行，稽查案件电子化管理已经得到普及，这不仅提高了税务稽查部门的工作效率，增加了税务稽查工作过程的透明性，减少了稽查人员违纪的可能性，同时数据库中也积累了大量的涉税违纪违法数据，各种违法违纪手段中存在着隐藏的规则，找出违法违纪手段之间存在的关联性可以帮助稽查人员在稽查工作中有目的地去查找相关账簿记录。如存在规则 M→N，则稽查人员在办案时发现存在 M 类型的行为，就应该重点检查是不是存在 N 类型的行为。

4 结果分析

挖掘出的 62 条关联规则符合税务稽查工作的实际，有一定指导意义，但原始稽查数据存在很大的“噪音数据”，影响了挖掘质量。其原因有以下几点：代码设计较粗，与基层实际操作有一定差距，也对实际数据录入产生影响；操作人员人为将多笔数据合成一笔数据，或随意选取任意一种违章手段输入，影响了原始数据的真实性和可利用性；预处理后，有效数据量相对较少，不能反映关联规则的全貌。

上面只是一些初步的稽查工作的分析、探讨工作，今后将收集更多的稽查数据以验证文中的挖掘算法，并将之用于相似业务的分析。

5 结束语

使用 WEKA 工具实现了关联规则挖掘过程的可视化、挖掘过程的可视化，用户可以根据需要适当调整所期望的最小支持度和最小可信度，以求能发现一些新奇的、反常的关联规则。随着数据仓库技术的发展，数据挖掘会越来越发挥其独到的分析优势，作为一种全新的解决问题的手段，数据挖掘带给稽查工作新的思考问题的角度和方式，必将会在未来的税务稽查工作中发挥其高效、智能的作用，为税务稽查的决策提供强有力的支撑。

参考文献：

[1] Witten L H, Frank E. 数据挖掘实用机器学习技术[M]. 第 2 版. 董 琳，邱 泉，于晓峰，等译. 北京：机械工业出版社，2006.

[2] 蒋丽华. 数据挖掘技术在税务稽查中的应用[J]. 税务研究，2007(5)：84－86.

[3] 范年茂，郭 冰. 电子商务对税务稽查的挑战[J]. 税收与企业，2003(8)：24－26.

[4] 董 涛，高 平，邹丁艳. 综合征管软件业务手册 v2.0 [R]. 济南：山东省国家税务局，2005.

[5] 崔贤岳，李际军. 数据挖掘技术在税务系统中的应用[J]. 计算机工程，2007，33(14)：283－284.

[6] 朱孝宇，王理冬，汪光阳. 一种改进的 Apriori 挖掘关联规则算法[J]. 计算机技术与发展，2006，16(12)：89－90.

[7] 胡 侃. 夏绍玮. 基于大型数据仓库的数据挖掘：研究综述[J]. 软件学报，1998，9(1)：53－63.

[8] Han Jiawei, Kamber M. 数据挖掘[M]. 范 明，孟小峰等译. 北京：机械工业出版社，2005.

[9] 杨健兵. 数据挖掘中关联规则的改进算法及其实现[J]. 微计算机信息，2006，7－3：195－197.

[10] 马盈仓. 挖掘关联规则中 Apriori 算法的改进[J]. 计算机应用与软件，2004，21(11)：82－83.

[11] 肖 智，李 勇，李昌隆. 一种基于相关分析的数据预处理方法[J]. 重庆大学学报，2002，25(6)：133－134.

[12] WEKA 中文站. WEKA 入门教程[EB/OL]. 2006－11. http://forum.wekacn.org/.