**湖 南 信 息 学 院**

**数据挖掘课程论文**

|  |  |
| --- | --- |
| **论文题目** | **数据挖掘在交通事故中的应用** |
| **学生姓名** | **邹琪** |
| **学号** | **201501420307** |
| **二级学院** | **电子信息学院** |
| **专业名称** | **软件工程** |
| **班级** | **15软件工程三班** |
| **指导教师** | **张钰莎** |

二〇一八 年 十 月 十九 日

**摘 要**

提出一种使用数据挖掘领域中的多维关联规则分析大量交通事故记录的方法,通过找出可能导致交通事故发生的频繁因素组合来发现某些事故发生的规律,并将这些规律作为现实中做出预防举措的依据。

**关键词**：交通事故；数据挖掘；关联规则；多维数据模型；Apriori改进算法

**ABSTRACT**

This article presents a method of ananlyzing a large number of traffic accidents' records by multidimensioned association rule in data-minning area. Through searching out the combination of frequent factors that might cause accidents, the regularity can be found as the foundation to the prevention in reality.

**Keywords:** Traffic Accident; Data Mining; Association Rule; Mutidimensioned Data Model; Apriori Improved Algorithm;

**目 录**

[**摘 要 II**](#_Toc527729646)

[**ABSTRACT III**](#_Toc527729647)

[**第一章 前 言 1**](#_Toc527729648)

[**第二章 系统的基本框架 1**](#_Toc527729649)

[**第三章 多维分析数据模型 1**](#_Toc527729650)

[**3.1 多维分析技术 1**](#_Toc527729651)

[**3.2 几个基本概念 1**](#_Toc527729652)

[**3.3 物理模型 2**](#_Toc527729653)

[**3.4 多维数据模型的特点 3**](#_Toc527729654)

[**第四章 关联规则挖掘技术 3**](#_Toc527729655)

[**4.1 关联规则的含义 3**](#_Toc527729656)

[**4.2 多维关联规则 4**](#_Toc527729657)

[**4.3 几个概念 4**](#_Toc527729658)

[**4.3.1 支持度（Support） 4**](#_Toc527729659)

[**4.3.2 置信度（Confidence） 4**](#_Toc527729660)

[**4.3.3 频繁项集 5**](#_Toc527729661)

[**4.4 多维关联规则的使用 5**](#_Toc527729662)

[**第五章 基于关联规则的Apriori算法实现及改进 5**](#_Toc527729663)

[**5.1 经典Apriori算法 5**](#_Toc527729664)

[**5.2 Apriori算法的改进思路 6**](#_Toc527729665)

[**第六章 结束语 7**](#_Toc527729666)

[**参考文献 8**](#_Toc527729667)

第一章 前 言

众所周知，交通事故的引发不仅和驾驶员的行为有关，而且与不同道路环境条件、天气条件、交通状况等有一定的联系，是综合因素共同作用的结果。如何从大量的交通事故数据中发现关联，分析事故的原因，以便于有关部门采取相应的对策，值得研究。将大量杂乱无章的数据组织为信息，从信息中进一步发现规律是本系统的基本思路。

第二章 系统的基本框架

借助借助数据挖掘领域中的多维关联规则技术对交通事故记录进行分析,原始数据可能存在错误,先通过清理大量原始记录行数据,并根据交通事故诱因特点提炼出与发生事故有联系的记录字段数据,组成新的事故记录表,从表中找出诱因记录字段和事故结果字段组成的频繁字段组合，最后通过预先给定的“支持度”和“置信度”来衡量该频繁字段组合是否能推导出事故关联规则。



第三章 多维分析数据模型

**3.1 多维分析技术**

交通事故数据和商务数据一样，本质是多维的。多维数据模型以描述描述分析数据的多维特征为目标，最终形成一个模拟现实的多维逻辑模型。在这一模型中，数据不再按实体和关系进行组织，而是以交通事故数据为例，时间时间维度由日、周、月和年组成，表示事故发生的时间；位置维度可以由公路的里程桩号桩号表示，表示事故发生的地点；交通事故性质维度由一般事故、重大事故和特大事故组成，表示事故的严重程度；天气维度表示事故发生时的天气状况等等。需要指出的是各维度的取值可以是连续值，也可以是间断的个别点。

**3.2 几个基本概念**

1. 层次。维的层次关系是根据对数据组织详细的要求不同来建立的，以达到从不

同层次对数据的聚合和分解。

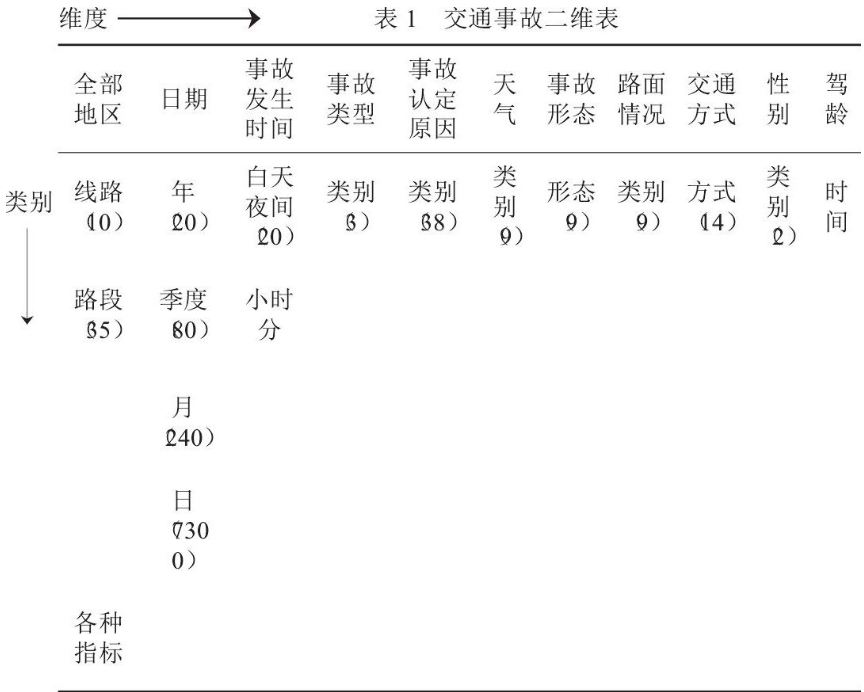
1. 类别（详细类别）。类别是组织多维数据的另一种方法。按事故的性质，可分为一般事故、重大事故的特大事故三种类别；按事故形态可以分为正面相撞、侧面相撞、尾随相撞等若干类别。各维度的类别数目是有限的。
2. 指标量。指标量是在维度空间衡量事故信息的一种方法，是对事故数据进行定量的衡量，如亿车公里事故数、亿车公里死亡率。

**3.3 物理模型**

采用关系型数据库的二维表表达多维概念时，可以将多维空间结构划分为两类表，一类是事实表，用来存储指标实体的度量值和各个维度的关键字；另一类表是维表，是维的描述信息，包括维的层次以及类别。由于交通事故数据的累积相对来讲是海量的，在具体建立物理模型中必须对每个实体的数据容量和更新频率进行评价，以便于分析工作的顺利开展。通常事实表是大容量的，而维表是小容量的，事实表的更新频率远远高于维表。

表1中，第一行表示各个维度，每一列表示不同唯独的类别，其中括号内的数字表示各类别的数目。通常类别和一个维度的关系应保持一定的比率，如1：10，这将有助于用户检索和理解数据，提高数据的可利用性。

在指标栏里，定义了三种指标量：年亿车公里事故数、年直接经济损失和年亿车公里死亡数，以这三个指标为中心展开分析。



**3.4 多维数据模型的特点**

为下一步的多维关联规则挖掘提供了基础。从概率论的角度考虑，交通事故属于小概率事件，事故的发生有其偶然性，但也有一定的必然性，存在着一定的规律性。所谓数据挖掘时采用机器学习、统计等方法进行知识学习的阶段，是知识发现（KDD）的核心。数据挖掘的任务是从数据中发现模式，按功能模式可分为两大类：预测型（predictive）模式和描述型（descriptive）模式。在实际应用中往往根据模式的实际作用细分为：分类模式、回归模式、时间序列模式、聚类模式、关联模式、序列模式。一般在建立这些模式时，使用一部分数据作为样本，用另一部分数据来检验、校正模式。

第四章 关联规则挖掘技术

**4.1 关联规则的含义**

简单地说数据挖掘中的关联规则技术能推导出一种“A=>B”的模式，即当A事件发生的时候，B事件会伴随发生，称为一种A与B的关联关系。例如，可以将A事件定义为“路滑”，而将B定义为“撞车”，那么A=>B不难理解为当路滑的时候会发生撞车事故。

**4.2 多维关联规则**

按照多维数据库的术语，蕴涵式中的每个不同的谓词都称作维。例如对于某商场的事务数据库来说，由谓词buys参与的关联规则表述如下：

buys （X，"IBM desktop computer"）=>buys （X，"Sony print-er"），

其中X是变量，其后跟的引号部分内容为其值。这种关联规则称为单维关联规则。若引进新的谓词age和occupation，并使用

age （X,"20 ~29"）^occupation （X,"student"）=>buys （X,"note-book PC"）

其中X是变量，其后跟的引号部分内容为其值（含义为：年龄在20至29岁的学生倾向于买笔记本电脑）。以上的推导式涉及多个谓词，就是多维关联规则。

**4.3 几个概念**

**4.3.1 支持度（Support）**

以上关联规则所表述的是一种关联关系，并且这种关联关系是在一定概率约束下才有意义的。拿开始的A=>B例子来说，如果A，B这一事件组合在总事务记录中出现概率过低，低于预先给定的临界值，则说明这一事件组合发生的次数太少，少到我们没有必要关注它的程度。相反，如果这一事件组合的出现概率高于预先给定的临界值，说明这个模式具有一定程度的普遍意义，我们发现并进一步研究它是有价值的。我们称以上的这个概率为“支持度”。这里的临界值的设置，要根据领域专家的建议和实际情况来综合决定。

**4.3.2 置信度（Confidence）**

在进一步研究A=>B这一模式的过程中，我们有必要考虑另一个概率约束，称为“置信度”。考虑到即使A，B这一事件组合发生频繁，但如果这一事件组合的发生次数比起A事件的所有发生次数来说，其值过低，低于预先给定的临界值的话，说明A的发生不足以导致B的发生。使用以上例子做进一步阐述：假设平均在100条事故记录中都存在“路滑”这一因素，但在相应的100条事故记录中只有一条存在“撞车”这一结果，这时，A事件的发生导致B事件发生的概率平均来说只有1%，这不足以支撑A能推导出B这一结果。这里的临界值的设置同样也要根据领域专家的建议和实际情况来综合决定。

**4.3.3 频繁项集**

如果用户设定一个最小支持度为min\_sup，那么那些支持度大于min\_sup的项集被称为频繁项集。

有了上述几个概念之后，我们可以对关联规则作一些详细的描述。所谓关联规则，就是一个形如A=>B的逻辑蕴涵式，其中A∈T，B∈T且A∩B=Ф。当事务数据库中有S% 的事务包含A∪B，则关联规则的支持度可以被定义为：

Support（A=>B）= Support （X U Y）=S

若事务数据库D中包含A 的事务中有C% 同时包含Y、则关联在规则A=>B的可信度Confidence可定义为：

Confidence（A U B）/Support（A）

一般认为，一个关联规则具有形式A=>B,其中A和B都是项集，对于这个规则的解释就是：如果其中一个事务A出现，则事务B也可能同时出现。

**4.4 多维关联规则的使用**

在将多维关联规则应用于交通事故记录数据库时，需要首先使用数据清理技术挑选出我们需要的记录字段，然后使用合适的算法挑选出频繁字段值组合，最后推导出满足预先给定的“支持度”和“置信度”的值的关联模式。并将这些模式作为我们分析当前状况和采取行动避免交通事故的依据。

同时满足最小支持度阈值和最小置信度阈值的规则称为强规则。挖掘关联规则问题就是产生支持度和可信度分别大于用户给定的最小支持度和最小置信度的关联规则，也就是产生强规则的问题。

第五章 基于关联规则的Apriori算法实现及改进

Apriori算法是一种最有影响的挖掘布尔关联规则频繁项集的算法。因为本文描述的交通事故记录中要求得的是一种因果的关系，符合布尔关联规则的适用情况，所以使用Apriori算法是合适的。

**5.1 经典Apriori算法**

Apriori算法最主要的概念，就是从候选项目集合中藉着扫描一次资料库，找出大于或等于使用者所定义的最小支持度者，称之为常见项目集合；再以常见项目集合（1argek k item-sets）组成候选项目集合（candidate k+l itemsets）。如此重复的运算直到无法找到常见项目集合为止。

经典Apriori算法的性能分析：

（1）经典的Apriori算法在每次生成常见项目集合时，都会扫描一次完整的资料库。随着数据库容量的增大，重复访问数据库、外存的这种扫描将会大大增加I/O开销，并且这种开销将随着数据库中记录的增加呈现出几何级数的增加,将导致整个系统性能低下。

（2）由常见项目集合生成候选项目集合时，会产生庞大候选集合。

因此经典的Apriori算法时间和空间复杂度都比较高，不太适合数量非常大的交通事故分析。

**5.2 Apriori算法的改进思路**

与经典Apriori算法不同，改进的算法在产生k- 1维频繁项目集时，可以同时实现对该集中出现元素的个数进行计数处理。这些数据将会为随后的删除无效项集提供依据。例如，某一个元素要成为K维项目集的一个元素的话，该元素在k- 1阶频繁项目集中的计数次数必须达到k- 1个，否则不可能生成K维项目集。因此，对某元素而言，若它的计数个数不到k- 1的话，可以事先删除该元素，从而排除由该元素将引起的大规格的所有组合。然后再按Apriori算法检验新的K维频繁项目集的所有k- 1维项目子集是否包含在已经求出的k- 1维频繁项目集内。如果其中有一个没有包含，则也可删去该k维项目集。这样得到的K维频繁项目集就是一个真正有用的项目集合。得到这个项集以后，对数据库D进行重新扫描。数据库中每条记录都有一个标记。记录着这条记录是否属于删除队列。起始状态每条记录都标识为“保留”，从头开始按顺序扫描数据库。如果某事务不包含Ck中任一项集，则将这条记录打上“删除”标签，并将其与数据库末端第一条标记为“保留”的记录交换位置，把新交换上来的记录作为下一条要扫描的对象。否则保留该记录，直接扫描下一条，直到整个数据库的记录全部被扫描完，得到新的数据库D 。下一次扫描数据库的时候，就只需要扫描D中标记为“保留”的记录就可以了。这样，随着K的增大，D中的事务量就可以大大减少，对于下一次扫描可以大大节约I /O开销。

改进算法较经典Apriori算法有明显的优点：首先由于它在每次生成候选项集之后都会对其进行有效的筛选，删除其中没有用的项集，因此可以大大减小了下一步接连生成的相集规模。其次，由于它不仅有效筛选项集，同时也会有效地减少数据库中记录的数量，对于像Apfiof这样需要重复扫描数据库的算法而言，数据库中记录大规模地减少，也对提高算法运行所需要的时间和空间起到很大的作用。数据库的改写无疑也需要增加I /O消耗，但是，随着数据库规模的增大和算法循环次数的增加，这种由于“新生成的数据库”中记录的减少而使得扫描、比较时间的大大减少，其优势将会很快体现出来。

第六章 结束语

面对日益严峻的道路交通安全形势，交通管理部门越来越重视对交通事故数据的收集和分析工作。在分析道路交通事故现状的基础上，应用数据挖掘技术，可以更为完善的处理复杂、稀疏、多维、不全的数据，从而做出更为科学的预测。因而，应用数据挖掘技术分析预防道路交通事故，对于保障人们的人身安全，减少国家的经济损失有着积极作用。相信数据挖掘技术会为道路交通事故分析预防工作提出新的思维方式。

在交通事故发生的监管中，利用数据挖掘技术可以从多个方面去预防和预测交通事故；例如可以使用数据挖掘中的聚类分析，预测那些人发生交通事故的的几率大，从而可以指定某些政策，使得具有某些特征的人限制驾驶汽车上路；还可以使用数据挖掘技术，预测某些地段是常发生事故的地区，可以对该地进行改造，或者提示司机注意事故多发地，这一应用已经在现实中的到了应用；利用数据挖掘，对发生交通事故的所有司机曾今考取驾照所在的驾校进行统计，对在某驾校考取驾照的司机而又发生的交通事故较多的驾校进行严格的审查，是否有不合规的人考取了驾照。

随着汽车数量的不断增加，道路也越来越复杂，我们遇到的交通安全问题，将会越来越复杂，数据挖掘的应用也越来越广，效益也非常的显著，因此可以在交通安全问题中使用数据挖掘技术，可以对交通安全问题进行预测，进而达到未雨绸缪的功效，为社会带来更安全的出行方式。

参考文献

[1]陈玲燕. 出租车GPS数据的信息挖掘研究[D].北京交通大学,2018.

[2]杜娜娜. 基于GPS数据的交通行为及出租车行驶路线推荐研究[D].北京邮电大学,2018.

[3]洪爱斌.道路交通安全评价中数据挖掘技术的应用分析[J].建材与装饰,2018(18):254-255.

[4]吴紫恒. 基于数据挖掘的驾驶行为与道路拥堵分析[D].中国科学技术大学,2018.

[5]栾文波,汪林,张宁,赵圣娜.基于数据聚类技术的城市轨道交通运营特征日分类研究[J].城市轨道交通研究,2018,21(04):14-17.

[6] 杨希刚.数据挖掘在交通事故中的应用[J].软件导刊,2008(06):18-20.