文章编号:1671-8879(2015)05-0097-07

高速公路收费数据中环境-运营特征关联规则挖掘

杜 瑾[1],郝 珺[2],樊海玮[1]
(1. 长安大学 信息工程学院,陕西 西安 710064; 2. 西安铁路局,陕西 西安 710054)

摘 要:针对传统关联规则挖掘算法在海量高速公路收费数据挖掘交通规律时存在的运算冗余、复杂度高、代价大等问题,提出一种改进的 Apriori 算法,可获得收费数据中环境特征与运营特征间的关联规则。首先将收费数据特征属性划分为收费站信息、时间信息、气象信息、运营信息 4 类,并进一步划分为环境特征域与运营特征域,从而构建收费数据特征模型;其次提出基于目标域的关联规则挖掘算法,在传统 Apriori 算法频繁项集产生过程中,根据期望规则结构以及已知规则后件,过滤无效项集,形成仅包含目标域的关联规则;最后,讨论了高速公路收费数据挖掘中的数据清理、离散化与数据拟合等问题,并在西宝(西安—宝鸡)高速公路收费数据集上进行了验证。试验结果表明:提出的方法时、空代价优于传统算法,能有效减少频繁项集数量,规则置信度在 93% 以上;试验中得到收费站位置、通行时间、气温及是否免费日等环境特性对车辆运行速度和收费站通行率的影响规则,这些规则可为高速公路交通控制及信息服务提供支持。

关键词:交通工程;交通信息;数据挖掘;关联规则;目标域;高速公路收费数据;运营特征
中图分类号:U491 **文献标志码**:A

Association rules mining between environment and running characters in expressway toll data

DU Jin[1], HAO Jun[2], FAN Hai-wei[1]
(1. School of Information Engineering, Chang'an University, Xi'an 710064, Shaanxi, China;
2. Xi'an Railway Bureau, Xi'an 710054, Shaanxi, China)

Abstract: Traditional association rule mining algorithms do not distinguish the data characters in expressway toll data mining, which may cause some problems such as calculation redundancy, high complexity and high cost. This paper proposed an improved Apriori algorithm, which could obtain association rules between environment characters and running characters of toll data. Firstly, the toll data characters were divided into four classes including toll station information, time information, weather information, and running information. Further, those characters were divided into environment domain and running domain. On this basis, the toll data model was built. Secondly, this paper proposed an association rule mining algorithm based on OD (objects domain). In the generation of frequent item set of traditional Apriori algorithm, the invalid item

收稿日期:2015-05-10
基金项目:国家自然科学基金项目(51278058);陕西省交通厅科技项目(13-39X);中央高校基本科研业务费专项资金项目(CHD2011JC02)
作者简介:杜 瑾(1974-),男,陕西西安人,讲师,工学博士,E-mail:dujin@chd.edu.cn。

sets were removed according to expected rules' structure and known rules' consequents, and association rules containing objects domain item were retained. Lastly, various issues such as data cleaning, data discretizing, and data fitting, were discussed. This algorithm has been tested in association rule mining of expressway tolling data for Xi'an-Baoji expressway. Experiment results indicate that this algorithm is better than traditional ones on time space cost, and can reduce the amount of invalid frequent item set efficiently. The confidence of rules is over 93%. The association rules generating from experiment demonstrate the effects of environment characters such as toll station location, passing date and time, temperature and free day on vehicle speed and passing rate of toll station. Those rules would support expressway traffic management and information service. 5 tabs, 15 refs.

Key words: traffic engineering; traffic information; data mining; association rule; expectation domain; expressway toll data; running character

0 引 言

截止到 2014 年 12 月底,中国高速公路通车总里程数已达到 111 918 km[1]。目前,中国高速公路建设增长速度已趋于稳定,然而,如何有效提高当前高速公路的通行效率,成为当前交通运输领域研究的主要方向之一[2]。在不投入新的建设资金的情况下,通过交通信息化管理与服务,最大限度地发挥高速公路的通行效用,这需要公路建设、公路检测、公路运输管理以及信息技术等多学科技术的融合,为高速公路运营管理提供有效、科学的信息服务与决策支持。

有关高速公路运营方面的数据类别众多,例如公路建设部门部署的公路质量监控传感数据,或交管部门部署的交通传感或视频监控数据,但是其中数据量最大的还是高速公路收费站的收费数据。据统计,中国各个省每年的高速路网通行车辆均超过亿辆·次。收费数据为高速公路车辆行驶及收费站运营情况分析提供了海量的数据平台,对其进行深入的数据挖掘和分析,可以找出许多潜在的高速公路交通运营规律,从而为高速公路信息化建设提供有利的支持[3]。高速公路收费数据挖掘研究主要是利用聚类方法对收费数据进行特征分类,从而发现运行特征分布或者异常数据[4-7],然而在关联规则挖掘方面研究较少,不能有效发现特征之间的关联关系。中国郑长江等利用 Weka 软件对无锡收费站数据进行挖掘,获得收费时间、进站车速、收费车辆数、收费站规模方面的关联关系,但其研究只是利用现有的数据挖掘工具,对单一收费站数据进行分析,涉及数据量少,存在挖掘效率不高,分析结论较为片面等问题[8]。日本早稻田大学在面向交通控制分析的关联规则挖掘方面多有建树,其中 Zhou 等利用遗传网络设计,挖掘时间相关的频繁交通状态序列模式,从而对交通流量趋势进行预测,然而序列模式挖掘虽然可以发现数据特征值次序上的依赖关系,但不同数据特征之间的关系却无法发现,尤其是不能形成"A→B"规则形式的交通规律[9-11]。此外,常规的挖掘过程没有结合高速公路收费数据特点,其过程是先进行挖掘计算,再通过模式评估删除那些对研究无用的规则。然而这些无用规则在挖掘计算中却消耗了大量的时间和空间资源。为此,本文以高速公路收费数据为研究对象,采用关联规则挖掘,从而获取收费站基本信息、运营信息,时间信息以及气象信息等不同属性域内众多属性之间多对多的关联关系。结合研究目标这种关联规则前、后件分属不同属性域的特性及频繁项集的单调特性,本文对传统的 Apriori 算法进行改进,获得以目标属性为规则后件的交通运营规则。这些规则具有较高的针对性,在提升计算效率的同时,可为高速公路交通诱导,公众出行信息服务提供有效的决策支持。

1 高速公路收费数据建模

高速公路收费数据具有规模大、属性多、表结构复杂等特征。以陕西省高速公路收费数据为例,在数据量方面,2013 年全省收费数据共 1.5 亿条记录,文件存储超过 60 G。在数据表示和组织方面,存在多个与高速公路运营相关的数据库表,这些表属性众多,相互关联,可进行深入的挖掘分析。主要涉及的数据库表及关键字段如下页表 1 所示。

除了上述数据库表属性外,还可通过计算与数据拟合的方式,获得更深层次的推导属性,例如车辆的平均通行速度、收费站出入通行交通量、当天分时

温度等。这些属性进行有机组织、加工、转换后，可构建成适用于研究目标挖掘的数据集。

表 1　高速公路收费数据库表内容

Tab. 1　Content of expressway tolling database

数据库表	包含内容
车辆出口收费数据表	入口站编号，入口日期及时间，出口站编号，出口日期及时间，终判车型，终判车种，总费额(元)，付费方式，车辆牌照，付费细分种类，费率表版本，拆分金额……
收费费率数据表	编号，入口站编号，出口站编号，车型，车种，费率，两站距离
收费站信息表	编号，收费站编号，收费站名称，所属高速公路编号，所属高速公路名称，所属地区
气象信息	编号，日期，天气情况，最高温度，最低温度，风力情况，所属地区(市县)

高速公路收费数据虽然属性众多，但是可以将其划分为两类：一类为环境特征属性，主要描述收费站静态特征或不可控自然特征；另一类为运营特征属性，这类属性描述了有关收费站运营或车辆运行的动态变化特性，可通过交通控制手段进行调整。本文的研究目标旨在通过关联规则挖掘，发现环境特征对运营特征的影响关系及程度，从而为高速公路交通控制提供信息服务支持。

本文通过数据分析，建立高速公路收费数据模型 M，对数据特征进行形式化描述。M 可由以下元组表示为

$$M = < S, T, W, F > \tag{1}$$

式中：S 为收费站信息，包括收费站 ID、等级、车道数、所属高速公路、所属地区等属性；T 为时间信息，包括日期、时刻、是否法定节假日、是否免通行费日等属性；W 为气象信息，包括地区、天气(如晴、多云、小雨、霾等)、风力、分时温度等属性；F 为收费站运营信息，包括单位时间出口流量、入口流量、入口平均速度、出口平均速度等属性。

式(1) 中，S、T、W 可以看作是收费站环境特征；F 则看作收费站运营特征。

本文在上述数据平台之上，提出一种带目标域的改进 Apriori 算法。以下介绍关联规则挖掘的相关概念、问题描述、算法流程、算法性能分析以及试验分析结果。

2　相关概念及问题描述

本文对关联规则相关概念及形式化描述进行定义[12]：

项集(Item Set)：$I_S = \{i_1, i_2, \cdots, i_m\}$，其中 i_j 称为项；

事务 R，是项集的一个子集，$R \subseteq I$，每个事务 R 都有唯一表识 R_{ID}；

数据集 D：为所有事务集合；

包含：X 为项集 I 的一个子集，如果 $X \subseteq R$，则称 R 包含 X；

规则形式：前件 → 后件[支持度，置信度]，其中前件、后件分别为项集；

关联规则：$X \rightarrow Y, X \subset R, Y \subset R, X \cap Y = \varnothing$

支持度：事务集中即包含 X 也包含 Y 的事务的概率；

置信度：项集 X 出现在事务集 D 中，同时项集 Y 也出现的概率。

将收费站环境特征(收费站信息、时间信息与气象信息) 与收费运营特征关联关系发现问题转化为关联规则挖掘问题，其问题描述如下：

属性域：高速公路收费数据可由属性集 I 描述，I 可分割为 4 个子集：$I = I_S \cup I_T \cup I_W \cup I_F$，且任意 2 个子集相交为空集，即 $I_i \cap I_j = \varnothing$，其中 $i \neq j$，$i, j \in (S, T, W, F)$。则称 I_i 为一个域。域 I_i 为某个高速收费特征属性域，可表示为 $I_i = \{i_1, i_2, \cdots, i_j\}$；其中 i_j 为属性域 I_i 中第 j 个属性；

项：高速公路收费数据模型中的某个带值属性可当作一个项，它描述了某收费站 S_x 在特定时间段 T_z 以及特定气象环境 W_z 下的收费站运营情况 F_z，其中的某一条属性具体表现，可表示为 $i^z_{S_x} = \text{value}(S_x, a_z)$，其中 $a_z \in I$；

项集：收费站集 S 在特定时间段 T 以及气象环境 W 下的收费站运营情况 F 所有属性所构成的值域即为项集，可表示为

$$I_S = \{x \mid \forall S_I, \forall a_j, x = \text{value}(S_i, a_j), S_i \in S \wedge a_j \in I\} \tag{2}$$

事务集可由所有收费站特征集表示，是收费站基本信息库、时间信息库、气象信息库与收费站运营信息库的联合；

事务 R 则是对某收费站 S_i 在特定时间段 T_j 运营情况的全面描述，其收费站 ID 和时间戳联合可看作是事务 ID。通过事务 ID，将该收费站在某时间段内的所有特征向量(包括收费站情况特征、时间特征、气象特征、运营情况特征) 进行联合构成了一条特征向量，特征向量里的所有属性都可看作是无差异的；

期望的关联规则格式 $X \rightarrow Y$，其中 X 与 Y 分别属于不同的属性域。

设置最小支持度和最小置信度作为特定阈值，以此获得收费站环境特征与运营特征间的关联规则，从而发现环境与运营之间的推导关系。

传统的关联规则挖掘算法，如 Apriori 或 AprioriTid，对项集内的所有项是不加区分的[13]。因此用这些算法对事务集做关联规则挖掘就可能会产生 3 种类型的关联规则模式：①$A \rightarrow A$；②$B \rightarrow B$；③混合型。尤其是最后一种规则，其前件和后件既有可能包含收费站环境属性，也有可能包含收费站运营属性。为了能获得特定格式的规则，这些算法会扫描挖掘出结果中的所有规则模式，再通过人工评估删除掉那些不满足格式要求的规则。这样，在分析和筛选中，就会产生一些不必要的计算代价和空间代价。但是如果结合具体研究领域和需求，尤其是当规则模式结构已知，或只关注包含特定特征属性的规则时，就可以在挖掘过程中进行识别和处理，从而达到减少频繁项集的数量和长度，提高挖掘效率的目的。鉴于此，本文提出改进的关联规则挖掘算法。

3 带目标域的改进关联规则算法

关联规则挖掘存在如下事实：

如果规则 $a_1,\cdots,a_i \rightarrow b_1,\cdots,b_j$ 存在，则 $a_1,\cdots,a_i \rightarrow b_1, a_1,\cdots,a_i \rightarrow b_2,\cdots,a_1,\cdots,a_i \rightarrow b_j$ 一定也存在。也就是说，任何频繁项集的子集，也一定是频繁的。因此，可以将问题 $I_A \Rightarrow I_B$ 转换为规则集形如 $\{\wedge a_i' \Rightarrow b_j'\}$ 的问题。

定义：目标域(object domain，OD)

假设规则的结构格式和规则的后件是预先知道的，期望得到格式形如 $b_i,b_j,\cdots,b_m \rightarrow f$ 的关联规则，其中 $\{b_i,b_j,\cdots,b_m,f\}$ 为属性集合，$b_i,b_j,\cdots,b_m$ 属于域 $I_B = I_S \cup I_T \cup I_W$，$f$ 属于域 I_F，规则后件 f 是已知，则称 f 为目标域。

同时，注意到如下事实：

定理：在带有目标域关联规则挖掘时，如果 k-项集 $\{b_i,b_j,\cdots,b_{k-1},f\}$(属性元组的长度为 k) 不是频繁项集，根据频繁项集的“单调”特性(一个频繁项集的子集也必然是频繁的)，则 $b_1,b_2,\cdots,b_{k-1}$ 是“无效”的频繁项集，不可能产生形如 $b_1,b_2,\cdots,b_{k-1},\cdots,b_n \Rightarrow f$ 的关联规则。

在该定理中，“无效”的含义为：即使 $b_1,b_2,\cdots,b_{k-1},f$ 可以产生频繁项集 $b_1,b_2,\cdots,b_{k-1},\cdots,b_n$，它也不能产生频繁项集 $b_1,b_2,\cdots,b_{k-1},\cdots,b_n,f$。

因此，挖掘过程中，从所有的频繁项集中删除形如 $b_a,b_b,\cdots,b_i,b_j,\cdots,b_n$ 的频繁项集，因为它只能产生形如 $b_a,b_b,\cdots,b_i \Rightarrow b_j,\cdots,b_n$ 的频繁项集。

因此，针对期望获得形如“环境特征 → 运营特征”这类格式的关联规则，本文提出一种“带目标域的关联规则挖掘算法”，该算法在保证关联关系挖掘具有正确性和完备性的同时，可有效提高算法效率，避免不必要的计算代价。

算法流程如表 2 所示。

从算法可以看出，在 k- 频繁项集处理过程中虽然增加了无效频繁项集的过滤删除工作，但对 $(k+1)$- 候选项集数量的减少却具有很大贡献。

这里，本文对改进的带目标域的关联规则挖掘算法和传统的 Apriori 算法进行时间复杂度的比较，从而评估改进算法的性能和优越性。

表 2 带目标域的改进 Apriori 算法流程

Tab. 2 Procedure of modified Apriori algorithm based on objects domain

输入：事务集 R
输出：所有符合目标域规则结构的频繁项集
挖掘流程：
Step1. $k=1$，扫描事务集 R，生成 k 阶频繁项集 L_k
Step2. 将 L_k(k 阶频繁项集)分为包含目标域的 k 阶频繁项集 L_{k1} 和不包含目标域的 k 阶频繁项集 L_{k2}；
Step3. 由 L_{k1}、L_{k2} 生成包含目标域的 $k+1$ 阶候选集 $C(k+1)1$；
Step4. 首先对 $C(k+1)1$ 中的项计数，生成包含目标域的 $(k+1)$-频繁项集 $L(k+1)1$；
Step5. 设包含在 $C(k+1)1$ 中而不包含在 $L(k+1)1$ 中的项为 $b_i,b_j,\cdots,b_k,f$；
Step6. 从 $Lk2$ 中删除包含 $b_i,b_j,\cdots,b_k$ 的项；
Step7. 由 L_{k2} 生成不包含目标域的 $k+1$ 阶候选集 $C(k+1)2$；
Step8. 对 $C(k+1)2$ 中的项计数，生成不包含目标域的 $(k+1)$-频繁项集 $L(k+1)2$；
Step9. $k=k+1$，重复 Step3～Step8，直到生成最大频繁项集。

由于带目标域的关联规则挖掘的复杂度计算和具体的事务序列有关，为说明方便，设关联规则挖掘中在生成 k 阶频繁项集 L_k 时，频繁项集的数目为 M_k，L_k 中包含的无效频繁项数目为 N_k。

传统的 Apriori 算法在生成 $(k+1)$ 阶候选项集 C_{k+1} 时的时间复杂度为 $C_{M_k}^2$；带目标域的关联规则挖掘算法在生成 $(k+1)$ 阶候选项集 C_{k+1} 时的时间复杂度为 $C_{M_k-N_k}^2$。

显然，当 $N_k \geqslant 1$ 时，后者与前者的时间复杂度之比为

$$C_{M_k-N_k}^2 / C_{M_k}^2 \approx (1-N_k/M_k)^2 \tag{3}$$

因此，带目标域的关联规则挖掘算法可有效地

减少 k 阶候选集 C_k 的数量，从而降低关联规则分析的时间复杂度。

4　试验与讨论

4.1　试验数据选择

由于高速公路收费数据的数据量巨大，为了避免在分析时间和空间上的无谓浪费，本文选择 1～4 月份西宝高速公路收费数据作为代表性数据进行研究。

其代表性之一为地理特性，西宝（西安—宝鸡）高速公路是中国“两纵两横”公路主骨架 G30 连云港—霍尔果斯国道主干线的重要组成部分，途经西安市、宝鸡市等 4 市 6 县（区）。沿途共有 17 个收费站，站间最长距离 156 km。

代表性之二为时间特性，1～4 月共有 120 d，历时冬春两季，其中法定节假日 40 d，高速公路免费日 12 d。

4.2　数据预处理

在进行数据挖掘之前，必须对挖掘数据进行必要的预处理工作，本文的预处理包括数据清理、数据离散化以及数据拟合 3 项工作。

4.2.1　数据清理

通过分析，发现现有收费数据中存在着大量的噪音数据，如表 3 所示。

数据显示从常兴到太白山 24 km 的距离仅用 1 s 就能到达，这显然不可能。这些噪音数据主要是因为收费计算机时钟不准确所导致，只要入口站或出口站一端计算机时钟错误，就造成在途时间过短或过长的情况。

由于西宝高速公路限速为 60～100 km/h，本文设定，若平均通行车速落在 30～160 km/h 范围之外，则视为异常数据，其原因可能是时钟错误或车辆非正常行驶（超速或故障处理）所致。经查询得知，异常数据共有 199 404 条，只占总数据（3 387 831 条）的 5.89%，因此本文采用抛弃的方式，一方面这些数据只占很小的比例，即使抛弃了这些数据，也不会影响数据挖掘的准确度，另一方面如果采用平均值修正法，往往会忽略各类车的通行特征，而人为导致特征分布趋于平均[14]。

4.2.2　离散化

离散化是为了获得更富于趋势性的规则，而对精确的数值进行区间划分的一种处理方法[15]。当区间划分的越粗，落在该区间的数据就越多，该区间出现的频度就越高，也就是关联规则挖掘中所谓的高支持度。但是如果区间划分过粗，就往往只能获得常见或已知的共性知识不能发现更加有意义的潜在特征。本文中需要进行离散化的数据主要包括：

（1）通行量

通过统计，通行量单位为车·次，本文根据通行量统计值的最高和最低值进行划分，平均分为 3 段。需要注意的是，由于收费站地理位置、等级不同，不同收费站的通行量区别会比较大，如果无区别进行划分，必然不能获得真正意义上的“空闲”“正常”“繁忙”等概念。因此离散化时必须针对每个收费站自身的历史收费数据单独划分通行量等级区间。

（2）车速

如前所述，设定正常平均车速为 30～160 km/h 范围之内，因此对车速离散化可按平均 5 段划分，设为低速、较低、中速、高速、超高速 5 个等级。

（3）气温

该属性是从 1 月至 4 月间采集，经历冬春两季，变化跨度达 40 ℃，因此该部分离散化采用平均 3 区间划分，分别为低温、中温和高温。

4.2.3　数据拟合

由于高速公路收费数据中不包含气象数据，因此气象特征只能从气象部门获得。然而，历史气象数据中仅提供每天的最高温度和最低温度，因此分时气温只能通过数据拟合的方式获得。根据常规，每天凌晨 5:00 时气温最低，下午 3:00 时气温最高。因此，分时温度根据以下公式计算可得，其中 $P_{\max}$ 为最高温度，$P_{\min}$ 为最低温度，P_t 为 t 时刻拟合出的温度

$$P_t=\begin{cases}P_{\max}-\dfrac{P_{\max}-P_{\min}}{14}(t+9), & 0\leqslant t<5\\ P_{\min}+\dfrac{P_{\max}-P_{\min}}{10}(t-5), & 5\leqslant t<15\\ P_{\max}-\dfrac{P_{\max}-P_{\min}}{14}(t-15), & 15\leqslant t<24\end{cases}\tag{4}$$

经数据清理、离散化、拟合处理后，构建出一个包含 17 个收费站 120 d 内 24 h 运营情况的收费样本数据集 D，共 48 960 条事务数据。

表 3　收费数据异常实例

Tab. 3　Exception example of toll data

车牌号	入口站	入口时间	出口站	出口时间	收费金额/元	站间距离/km	平均速度/(km·h^{-1})
陕 CK4＊＊3	常兴	2013-12-23,11:08:34	太白山	2013/12/23 11:08:35	10	24	86 400

4.3 试验过程

高速公路收费数据关联规则挖掘试验过程为:对数据集 D 运用带目标域的关联规则挖掘算法进行分析,其中设置支持度阈值由 0.3 向 0.05 递减,每次递减步长为 0.01。经模式评估分析,当阈值取 0.1 时可得到优化的"环境特征→运营特征"关联规则模式。

如表 4 所示,通过与传统 Apriori 算法的性能比较,试验证明:若规则具有特定的格式(前后件分别属于不同域)时,通过在挖掘过程进行项集筛选和过滤,有助于避免产生冗余关联规则,降低计算代价,提高挖掘效率,得到更为有趣的关联模式。

经过收费站环境特征与运营特征间的关联规则挖掘,可得两者之间的蕴涵关系,以下为挖掘得出的部分有效规则,如表 5 所示。

表 4 传统 Apriori 与带目标域 Apirori 算法比较

Tab. 4 Comparison between traditional Apriori and Apriori algorithm with objects domain

比较项	传统 Apriori 算法	带目标域 Apriori 算法
事务总数	48 960	48 960
挖掘规则总数	988	461
人工模式评估获得有用规则	133	133
挖掘时间/ms	14 368	11 880

表 5 基于目标域关联规则挖掘部分结果

Tab. 5 Some results of association rule mining based on objects domain

序号	规则	支持度	置信度
1	STARTH='低' CURRENTT='低' FREEDAY=0 ==>AVGTXL='低'	0.1	0.98
2	MEANSPEAD='低'==>AVGTXL='低'	0.1	0.93
3	STARTH='中午' CURRENTT='中等' FREEDAY=0 AVGTXL='中等' ==>MEANSPEAD='中速'	0.1	0.97
4	CITY=宝鸡 FREEDAY=0==>MEANSPEAD='中速'	0.1	0.93
⋮	⋮	⋮	⋮

表 5 所列规则含义为:①若凌晨上高速,气温低,非免费日,则通行量低;②若收费站出入车辆平均速度低,则收费站出入通行量低;③若中午上高速,温度中,非免费日,则收费站出入通行量中等;④若收费站在宝鸡市辖区,非免费日,则收费站出入车辆平均速度中等;……

由试验结果可以看出,虽然规则的支持度较低(占总数据的 10%),但规则的置信度较高(93%以上),也就是说,当规则中条件特征出现时,其结论特征出现的概率在 93%以上。因此通过上述方法获得的高速公路收费站环境特征与运营特征间关联规则对与后期的高速公路交通诱导以及公众出行信息服务研究奠定了基础。

5 结 语

(1)对高速公路收费数据进行建模,从收费站信息、时间信息、气象信息以及收费站运营信息 4 个方面对收费数据进行形式化描述,并将数据属性划分为环境特征域和运营特征域,提出期望的"环境特征→运营特征"关联规则结构。

(2)针对高速公路收费数据关联规则挖掘中规则结构特点,提出基于目标域的改进 Apriori 算法,在挖掘过程中识别并删除无效项集,减少候选项集个数,从而提高挖掘效率,减少无关规则数量。

(3)对原始数据进行数据预处理工作,提出噪音数据清理方法,数据离散化方法以及分时气温数据拟合方法,丰富了数据范围,提高了挖掘结果的有效性和有趣性。

(4)对西宝高速公路收费数据进行了试验分析,试验结果证明了本文方法的有效性和实用性,得出规则的置信度达 93%以上,可为高速公路信息服务提供有力支持。

参考文献:

References:

[1] 王春花. 浅析高速公路建设的财务管理[J]. 时代金融,2015(4):174-179.
WANG Chun-hua. Discussion on financial management of expressway construction[J]. Times Finance, 2015(4):174-179. (in Chinese)

[2] 高俊林,王树林. 公路交通信息化建设有关问题研究[J]. 中国交通信息化,2013(6):34-35.
GAO Jun-lin, WANG Shu-lin. Research on issues of road traffic informatization construction[J]. China ITS Journal,2013(6):34-35. (in Chinese)

[3] 杨聚芬,姜桂艳,李 琦. 基于收费数据的高速公路交通拥挤自动判别方法[J]. 哈尔滨工业大学学报,2014,46(12):108-113.
YANG Ju-fen, JIANG Gui-yan, LI Qi. The automatic traffic congestion identification of freeway based on charging data[J]. Journal of Harbin Institute of Technology,2014,46(12):108-113. (in Chinese)

[4] Li C S, Chen M C. A data mining based approach for travel time prediction in freeway with non-recurrent congestion[J]. Neuro Computing,2014,133(10):74-83.

[5] 黄志军. 高速公路联网收费系统防逃费研究与实现[D]. 长沙:中南大学,2009.

HUANG Zhi-jun. Research on anti-defraud of highway network toll and its implementation[D]. Changsha: Central South University, 2009. (in Chinese)

[6] 杨海陆. 公路收费系统数据分析与挖掘[D]. 哈尔滨：哈尔滨工程大学，2010.
YANG Hai-lu. Data mining and analyze in freeway charging system[D]. Harbin: Harbin Engineering University, 2010. (in Chinese)

[7] 徐晓帆，罗庆异. 数据挖掘技术在 ETC 运营管理中的应用[J]. 科技广场，2011(9)：81-85.
XU Xiao-fan, LUO Qing-yi. Application of data mining in ETC operations management[J]. Science Mosaic, 2011(9): 81-85. (in Chinese)

[8] 郑长江，沈金星. 数据挖掘在高速公路收费数据中的应用[J]. 中国科技论文在线，2008，3(10)：742-745.
ZHENG Chang-jiang, SHEN Jin-xing. Data mining applications to charge data of expressways[J]. Science Paper Online, 2008, 3(10): 742-745. (in Chinese)

[9] Zhou H Y, Shingo M B. Backward time related association rule mining with database rearrangement in traffic volume prediction[C]//IEEE. IEEE International Conference on Systems, Man, and Cybernetics. San Antonio: IEEE, 2009: 1047-1052.

[10] Zhou H Y, Shingo M B, Mainali M K, et al. Generalized association rules mining with multi-branches full-paths and its application to traffic volume prediction [C]//Fukuoka International Congress Center. ICROS-SICE International Joint Conference 2009. Fukuoka: SICE, 2009: 147-152.

[11] Zhou H Y, Mabu S, Li X N, et al. Generalized rule extraction and traffic prediction in the optimal route search[C]//IEEE. IEEE World Congress on Computational Intelligence, Barcelona: IEEE, 2010: 2625-2632.

[12] Agrawal R S R. Fast algorithm for mining association rules[C]//Morgam K. In Proceedings of 1994 International Conference on Very Large Databases. Santiago: DBLP, 1994: 487-499.

[13] Xiong C Q, Chen S B, Ding M. Association rules mining for discussion information in group deliberation support system[C]//IEEE. Proceedings of the 8th International Conference on Computer Science and Education. Colombo: IEEE, 2013: 67-70.

[14] 张峰伟. 一种 Web 使用挖掘数据清理方法[J]. 四川大学学报：工程科学版，2014，46(4)：160-165.
ZHANG Feng-wei. A novel data cleaning approach for web usage mining[J]. Journal of Sichuan University: Engineering Science Edition, 2014, 46(4): 160-165. (in Chinese)

[15] 张钰莎，蒋盛益. 连续属性离散化算法研究综述[J]. 计算机应用与软件，2014，31(8)：6-8.
ZHANG Yu-sha, JIANG Sheng-yi. Survey on continuous feature discretization algorithm [J]. Computer Applications and Software, 2014, 31(8): 6-8. (in Chinese)

(上接第 81 页)

[11] 牛　斌. 体外预应力混凝土梁弯曲性能分析[J]. 土木工程学报，1999，32(4)：37-44.
NIU Bin. The analysis of flexural behavior of externally prestressed concrete beams[J]. China Civil Engineering Journal, 1999, 32(4): 37-44. (in Chinese)

[12] 李红利. 节段施工体外预应力桥梁力学性能研究及面向对象的程序设计[D]. 长沙：长沙理工大学，2006.
LI Hong-li. Studies on mechanical behavior of segmental bridge with external tendon and development of OOP[D]. Changsha: Changsha University of Science & Technology, 2006. (in Chinese)

[13] 祝明桥，方　志，徐昌慧，等. 体外配筋预应力混凝土箱梁变形与裂缝试验研究[J]. 公路交通科技，2004，21(8)：42-44.
ZHU Ming-qiao, FANG Zhi, XU Chang-hui, et al. Experimental study on the deflection and crack of externally prestressed reinforced concrete thin-walled box girder[J]. Journal of Highway and Transportation Research and Development, 2004, 21(8): 42-44. (in Chinese)

[14] 徐　栋，魏　华. 体外预应力桥梁转向块结构分析及配筋研究[J]. 同济大学学报：自然科学版，2005，33(6)：722-726.
XU Dong, WEI Hua. Spatial analysis and reinforcement calculation by strut-and-tie model of deviators of external prestressing structure[J]. Journal of Tongji University: Natural Science, 2005, 33(6): 722-726. (in Chinese)

[15] 卢春玲，李传习. 体外预应力箱梁桥转向块配筋设计分析[J]. 中外公路，2008，28(6)：122-126.
LU Chun-ling, LI Chuan-xi. Design and analysis on reinforced deviators of external prestressed concrete box girder bridge[J]. Journal of China & Foreign Highway, 2008, 28(6): 122-126. (in Chinese)

[16] 徐　勋，强士中. 体外预应力混凝土梁承载全过程分析新模型[J]. 西南交通大学学报，2007，42(6)：732-738.
XU Xun, QIANG Shi-zhong. New model for full process analysis of bearing behavior of externally prestressed concrete beams[J]. Journal of Southwest Jiaotong University, 2007, 42(6): 732-738. (in Chinese)