第33卷 第17期 计 算 机 工 程 2007年9月
Vol.33 No.17 Computer Engineering September 2007
·软件技术与数据库· 文章编号：1000—3428(2007)17—0069—02 文献标识码：A 中图分类号：TP18

铁路货票数据的知识发现

周东北，雷定猷
（中南大学交通运输工程学院，长沙 410075）

摘 要：铁路货票蕴含着极为丰富的信息资源，它是铁路生产经营管理的重要基础数据。该文从3个方面对货票数据进行了知识挖掘。运用集合理论构造关系数据库特征关联模型，描述了特征规则知识的表达，提出了算法，并对实际的货票数据进行了知识挖掘，分析了知识对营销的启示；运用聚类知识挖掘模型，从货物运距、货物运价号和客户对公司的收入贡献等方面探讨了货运市场的细分，挖掘出来的知识明确了铁路货运目标市场的选择；运用ARIMA模型对货票数据进行了季节性知识挖掘，用1999年—2005年的历史数据估算2006年的货运量。

关键词：货票数据；知识发现；聚类分析；市场细分

Knowledge Discovery in Railway Freight Invoice

ZHOU Dong-bei, LEI Ding-you
(School of Traffic & Transport Engineering, Central South University, Changsha 410075)

【Abstract】There are rich information resources in the railway freight invoice. It is important and basic data for the prosecution and management of the railway production. This paper from the three aspects knowledge excavates the freight invoice. The characteristic relative model is developed and the characteristic rule knowledge expression is abstracted by applying set theory. It presents the algorithms, knowledge excavates the practical freight invoice and analyzes the inspiration of marketing; applies the clustering to knowledge excavate the model and does research for the subdivision of goods transportation market from such aspects as the transportation distance, the sort of goods, the number of price for goods transportation and the clients' incoming contributions to the corporation etc. The knowledge from excavating has specified the choice of object markets for the railway goods transportation. The freight invoice data is handled by ARIMA method to be seasonal knowledge excavated. The freight amount in 2006 is evaluated according to the historical data from 1999 to 2005.

【Key words】freight invoice; knowledge discovery; clustering analysis; subdivision of market

铁路建成了铁路运输管理信息系统(TMIS)等一大批管理信息系统，这些系统的使用给铁路带来了巨大的经济效益和社会效益，同时也为铁路经营管理积累了大量的基础数据。在铁路TMIS中央货票库中，保存了大量货票信息，反映各地区间的货物交流状况及其特点[1]，充分挖掘这些数据，发现其中规律性的知识，为铁路各管理部门提供决策支持，是一件很有意义的事情。

1 特征关联规则知识发现

1.1 关系数据库特征关联模型

按集合论的方法[2]，定义特征关联规则的模型如下：

定义1 设D是一个关系数据库，R表示D中所有记录的集合，$R=\{r_1,r_2,\ldots,r_n\}$，At是字段集合，$At=\{a_1,a_2,\ldots,a_m\}$，V是D中字段取值的集合，$V=\{v_{11},v_{12},\ldots,v_{mn}\}$。

定义2 设A为At的子集，f是a和r的函数，$v=f(a,r)$，对于$a\in A$，$r_i\in R$，$r_j\in R$，$i\neq j$且$f(a,r_i)=f(a,r_j)$，则称r_i,r_j基于字段集A等价，基于A的所有等价记录的集合称为基于字段集A的等价类。A中字段的取值称为类别。

定义3 设A为At的子集，对于 $a\in A,v\in V,r_i\in R,r_j\in R$, $i\neq j$，有$f(a,r_i)\oplus v$， $f(a,r_j)\oplus v$成立，则称r_i,r_j基于字段集A和运算⊕等价，基于A的所有关于运算⊕等价记录的集合称为基于字段集A和运算⊕等价类，其中⊕为布尔运算符=、≠、<、>、≤、≥。定义2是定义3中运算⊕取“=”号的特例。

定义4 设S_p是一个给定的阈值，$0\leq S_p\leq 1$，对于支持度$S\geq S_p$的等价类，称为强类，反之则称为弱类。可以将数据库中的记录分类，对于强类需要进一步分析这些分类中的特征。

定义5 设E是一个基于A的等价类，A^c是A关于At的补集，B是A^c的子集，基于字段集B和运算⊕的等价类T称为E分类中的特征域，B中各字段的取值$\{b_1,b_2,\ldots\}$称为E分类中的特征。设S_E是E中的记录数，S_T是T∩E的记录数，则称$C=(S_T/S_E)*100\%$为特征置信度。设C_p是一个给定的阈值，$0\leq C_p\leq 1$，特征置信度$C\geq C_p$的特征域，称为强特征域，反之则称为弱特征域，强特征域字段的取值称为强特征，弱特征域字段的取值称为弱特征。强类中的强特征往往是具有代表性的知识。强特征规则知识的表达为：rule(E,T,C_p)。其中，E：类别；T：特征，它由字段集B和它的取值$\{b_1,b_2,\ldots\}$通过运算⊕组成；C_p：特征置信度。

1.2 算法描述

rule(E,T,C_p)算法如下：

For all A ⊂ At Do
求所有等价类集合$E_1,\ldots,E_m$;
For i=1 To m Do
If $S_i\geq S_p$ Then
对运算⊕求所有特征域$T_1,T_2,\ldots,T_k$;

作者简介：周东北(1974－)，男，博士研究生，主研方向：交通运输信息化；雷定猷，教授、博士

收稿日期：2006-10-08 **E-mail：**ding@mail.csu.edu.cn

For j=1 Tok Do
if $C_i \geq C_p$ Then
(E,T,C_p)存结果库
Endif
Next j
Endif
Next i
Next

1.3 规则举例

应用上述算法对铁路某单位 2005 年度货票数据进行了数据挖掘，它共有 226 381 条有效记录，给定分类支持度的阈值 10%，特征置信度的 15%，挖掘的分类特征规则 780 条。如：

例 1 rule(本公司收入，500km 以下，18.3%)。

例 2 rule(集装箱，8 号运价号，23.4%)。

下列的规则是降低特征置信度(取 1%) 从柳州局 2005 年第 1 季度货票数据库中挖掘的一条有启示的规则。

例 3 rule(NX_{17BT}，拖拉机，1.5%)。

例 4 rule({柳州南，NX_{17BT}}，拖拉机，94.2%)。

为什么例 3 和例 4 的特征置信度相差这样大呢？其原因是：例 3 是针对整个柳州局的关联规则，例 4 限定发站为柳州南站。一般的普通平车标重 60t，长 13m，NX_{17BT}是集装箱专用平车，标重 61t，长 15.4m，多数情况下不装拖拉机，这是例 3 特征置信度低的原因；另一方面，柳州产的拖拉机拆去附件后长 4.8m，重约 6t，用一般平车只能一车装两台，而用NX_{17BT}可一车装 3 台，货主多出 1t的整车费率，节约了 1/3 的费用，货主用集装箱专用平车NX_{17BT+}装拖拉机，这就是例 4 特征置信度高的原因。由例 3 和例 4 挖掘的知识说明， 铁路多年来把标重作为计算整车费率的唯一因素的办法应作些调整。例如考虑车型，长度。

2 聚类分析知识发现

聚类分析的基本思想是在样品之间定义距离，在变量之间定义相似系数，距离或相似系数代表样品或变量的相似程度。聚类方法有许多种，本文采用自组织神经网络聚类方法。

2.1 自组织神经网络聚类算法

这种网络学习是一种竞争型的学习，具体过程如下：

(1)初始化。将连接权W_{ij}随机地赋以[0,1]区间的某个较小值，设置处理单元j的邻域初始半径$N_j(0)$，可随机取一大值。

(2)提供输入。设处理单元 i 的输入样本

$$X=(x_1,\ldots x_i,\ldots,x_n)$$

(3)计算距离。计算j单元的各输入x_i与连接权W_{ij}的欧式距离d_j：

$$d_j=\|X-W_j\|=\sqrt{\sum_{i=1}^{n}[x_i(t)-W_{ij}(t)]^2} \quad (1)$$

(4)选择最小距离对应的单元 j^*（j^* 又称为获胜单元）

$$d_{j^*}=\min_j d_j=\min_j\sum_{i=1}^{n}[x_i(t)-W_{ij}(t)]^2 \quad (2)$$

(5)校正权。在输出层内，在 j^* 周围半径 $N_{j^*}(t)$ 以内的邻域中，各单元的连接权值均要加以调整，调整的公式是

$$W_{ij}(t+1)=W_{ij}(t)+\eta(t)[x_i(t)-W_{ij}(t)] \quad i=1,2,\ldots,n \quad (3)$$

式中，$\eta(t)$ 是学习率，它随时间而衰减，一般定义

$$\eta(t)=\frac{1}{t} \quad 或$$

$$\eta(t)=0.2[1-t/10000]$$

$N_{j^*(t)}$ 一般为圆形邻域，其大小也是随时间收缩的；返回到第(2)步，直到满足$[x_i(t)-W_{ij}(t)]^2<\varepsilon$（$\varepsilon$ 为给定的误差)为止。最后确定获胜单元。

上面的第(2)步~第(5)步就是一种竞争学习。当在输入层提供输入向量 X 后，在输出层中寻找连接权向量与输入单元 i 最靠近的单元 j^*，此时 j^* 及其邻域中各单元被激活，而有输出 1，其它单元输出 0，即

$$y_i=\begin{cases}1 & j\text{为以}j^*\text{为中心的邻域}N_{j^*}(t)\text{内的单元时}\\ 0 & j\text{为其它单元时}\end{cases}$$

2.2 基于聚类的货运市场细分知识发现

所谓市场细分，就是按照购买者的需要、购买态度、购买实践等不同变量，把一个市场分为若干个不同的购买者群体的行为。

2.2.1 按客户对单位的本线收入细分

表 1 是铁路某单位 2005 年货票中的有关客户的数据，经过自组织神经网络聚类算法分析得出的结果。

表 1 铁路某单位 2005 年货票客户数据聚类分析结果

客户类型	客户数量	占总客户数比例/%	本线收入/元	贡献度/%
1	73	0.54	57 292 561	75
2	120	1.01	11 413 747	14
3	12 230	98.45	7 685 575	10

聚类分析结果对该单位货运营销按客户给公司的收入贡献细分市场有下列启示：占客户总数不到 2%的主要客户(1、2 类客户之和约 200 家左右)对该单位本线收入的贡献度超过 90%。客户营销管理要抓住主要矛盾，稳住主要客户，在此基础上再开辟新的业务。

2.2.2 按运距细分

运距是货票市场细分的一个重要的因素，货物运输距离大致可分为短途、中途和长途。三者之间的界限从来都是模糊的，各种运输方式均在不同的运距上有自己独特优势。表 2 是铁路某单位 2005 年货票数据各分段运距收入以及铁路总收入，有自组织神经网络聚类算法分析得出的结果。

表 2 铁路某单位 2005 年运距和收入聚类结果

运距/km	本线收入/元	总运费/元	聚类号
0~500	29 480 121	13 264 181	1
501~1 000	16 750 670	114 541 850	2
1001~1500	7 112 579	52 457 971	3
1 501~2 000	12 851 053	142 020 233	2
2 001~2 500	12 177 513	151 293 129	2
2 501~3 000	2 899 723	48 087 972	3
3 001~3 500	1 231 825	45 473 560	3
3 501 以上	1 083 720	4 213 155	3

聚类的情况分析 :过去认为 500km 内是公路的运输市场，500km~1 500km 是铁路的运输市场，照此表 2 的 1 001km~1 500km 内的运距应属于第 2 类，而不应是第 3 类。但从市场分析，不难发现其原因。运距在 1 000km 以上，有些物资已超出厂商辐射范围，造成货票量小，且这个运距很宜于重载汽车运输，公路也大量分流；至于 1 500km 以上，批量大、重型的货物仍会回归铁路，1 501km~3 000km 内仍属于第 2 类。

3 季节性知识发现

差分自回归滑动平均模型(auto regressive integrated moving average models, ARIMA)方法被誉作时间序列预测方法中最复杂最高级的模型。下面用 ARIMA 模型挖掘估算某货运站 2006 年货运量。

设货运量预报周期为 s，疏系数 ARIMA 模型的形式为

$$\varphi(B^s)\nabla_s^D X_t=\Theta(B^s)E_t$$

其中，$\varphi(B^s)=1-\varphi_1 B^s-\varphi_2 B^{2s}-\ldots-\varphi_p B^{ps}$；$\theta(B^s)=1-\Theta_1 B^s-\Theta_2 B^{2s}-\ldots-\Theta_q B^{qs}$；$X_t$为时间一时间系列。

（下转第 85 页）