杨隆浩,蔡芷铃,黄志鑫,等.出租车乘车概率预测的置信规则库推理方法[J].计算机科学与探索,2015,9(8):985-994.

ISSN 1673-9418 CODEN JKYTA8
Journal of Frontiers of Computer Science and Technology
1673-9418/2015/09(08)-0985-10
doi: 10.3778/j.issn.1673-9418.1409066

E-mail: fcst@vip.163.com
http://www.ceaj.org
Tel: +86-10-89056056

出租车乘车概率预测的置信规则库推理方法*

杨隆浩,蔡芷铃,黄志鑫,何 星,傅仰耿+
福州大学 数学与计算机科学学院,福州 350116

Belief Rule-Base Inference Methodology for Predicting Probability of Taking Taxi*

YANG Longhao, CAI Zhiling, HUANG Zhixin, HE Xing, FU Yanggeng+
College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350116, China
+ Corresponding author: E-mail: ygfu@qq.com

YANG Longhao, CAI Zhiling, HUANG Zhixin, et al. Belief rule-base inference methodology for predicting probability of taking taxi. Journal of Frontiers of Computer Science and Technology, 2015, 9(8): 985-994.

Abstract: Large scale of data, various types of low-level attributes and uncertainty of prediction information exist in probability prediction of taking taxi. To solve these problems, this paper offline deals with the GPS data of taxi and road network data by using mining algorithms in the large-scale trajectory data domain, then builds a belief rule-base by transforming various types of information with uncertainty into rules which are in form of the belief structure, after that uses RIMER (belief rule-base inference methodology using evidential reasoning) to get the final probability of any points on the road network. Finally, the GPS data of Beijing's taxi in November of 2012 are taken as an example to illustrate the usage of the online prediction method, and the results show the real-time and accuracy of the proposed method.

Key words: probability prediction; GPS data; road network data; belief rule-base; belief rule-base inference methodology using evidential reasoning (RIMER)

摘 要: 出租车乘车概率预测中存在数据量级大,底层属性类型多,预测信息不确定的问题。鉴于此,整合大规模轨迹数据范畴中现有的挖掘算法对出租车GPS数据和路网数据进行离线处理;将多类型的不确定性数据

* The National Natural Science Foundation of China under Grant Nos. 71371053, 61300026, 61300104 (国家自然科学基金); the Natural Science Foundation of Fujian Province of China under Grant No. 2015J01248 (福建省自然科学基金); the Science and Technology Project of Education Department of Fujian Province under Grant No. JA13036 (福建省教育厅科技项目); the Science and Technology Development Foundation of Fuzhou University under Grant No. 2014-XQ-26 (福州大学科技发展基金项目); the National Collegiate Innovation and Entrepreneurship Training Program under Grant No. 201310386030 (国家级大学生创新创业训练计划项目).
Received 2014-08, Accepted 2014-10.
CNKI网络优先出版:2014-10-17, http://www.cnki.net/kcms/doi/10.3778/j.issn.1673-9418.1409066.html

转换为具有置信结构的规则形式,并以此构建置信规则库;通过置信规则库推理方法(belief rule-base inference methodology using evidential reasoning,RIMER)在线预测路网道路上各个地点的乘车概率。以北京市2012年11月某天的出租车GPS数据为例说明该在线预测方法的应用。实验结果表明,该预测方法具有较高的实时性和准确性。

关键词:概率预测;GPS数据;路网数据;置信规则库;置信规则库推理方法(RIMER)

文献标志码:A **中图分类号:**TP18

1 引言

现代城市智能交通的主要特征是在城市交通运输系统中有效地运用通信技术、信息技术、系统集成技术和电子控制技术等建立起覆盖范围广、高准确度、高效率、实时的交通运输管理系统,使道路、驾驶员和车辆三者之间能够进行智能联系。借助该系统,车辆可以在道路上安全、自由地行驶。相关研究[1]表明,城市采用现代智能交通系统(intelligent transportation system)可以使得城市道路交通的通行能力提高2~3倍,使得交通拥挤率降低至原来的20%~80%,油料消耗较以前减少30%,停车次数可以降低30%,废气的排放可以减少26%,交通行车时间减少13%~45%,发生交通事故的可能性大大降低,有效地提高交通运输效率,进而给城市建设带来巨大的经济效益和社会效益。

随着智能交通系统的发展,出租车上已逐渐配备GPS设备,出租车GPS数据反映出租车的经纬度坐标、运行速度和车载状态等信息,通过对出租车GPS数据进行挖掘可为出行的乘客带来便利。目前国内外学者针对出租车GPS数据开展的研究主要有路段平均行程时间估计[2]、出租车乘客等待时间预测[3]、路段平均速度估计模型[4]等,而出租车乘车概率预测同样是出租车GPS数据挖掘领域中不可或缺的研究问题,其主要目的是通过对出租车GPS历史数据进行分析,计算不同地点在不同时间能打到车的可能性,以发现更容易找到空驶出租车的地点。现有方法[3]中主要采用预处理GPS历史数据得到停泊点的乘车概率,然后再推荐乘客到相对概率较大的停泊点乘车,而对于随意性较大的乘客而言,仅提供指定点的乘车概率很难满足实用性和准确性的要求。

本文针对出租车乘车概率预测中影响因素具有多类型、GPS数据不够精确的情形,提出了一种基于置信规则库推理[5]的出租车乘车概率在线预测方法。该方法首先需要对路网数据和出租车GPS数据进行离线处理;然后依此构建用于预测任意点乘车概率的置信规则库;最后依据询问点的经纬度和时间在线预测乘车概率。本文以北京市2012年11月某天出租车GPS数据为例,说明了该方法的应用和特点。

2 问题背景及模型假设

出租车乘车概率预测方法建立在对路网数据和出租车GPS数据处理的基础上,其具有4个显著问题:

(1)出租车每时每刻都在产生GPS数据,数据量级十分庞大,因此处理GPS数据需要花费大量的时间。

(2)使用路网数据表示道路时,道路本身是错综复杂的,在初始路网数据上进行建模十分繁琐,因此需要对路网数据进行膨胀细化处理,抽象出道路的“骨架”并保留其拓扑结构。

(3)收集到的出租车GPS数据属于时空间隔的离散点,因此需要与路网数据相匹配,以此表示出租车的行驶轨迹,最后再构建出用于出租车乘车概率预测的置信规则库。

(4)出租车乘车概率预测问题实质上是一个具有不确定信息的多属性决策问题,需要引入成熟的理论在路网数据和出租车GPS数据的基础上对乘车概率进行准确的推算。

上述前3个问题在轨迹数据检索与挖掘领域已得到充分的研究与发展,并获得了许多有价值的成果[2-4],而对于第4个问题,本文拟引入置信规则库推理方法。该方法是杨剑波等人[5]在2006年提出的一种处理不确定性的理论和方法,其以扩展的IF-THEN规则作为知识库,以证据推理算法作为推理机,因此能在不需要先验概率的前提下,用简单的推理形式

得出理想的结果，为不确定性信息的表示及处理提供了有效的方法，目前已成功应用于信息合成和不确定推理等领域[6-8]。

对于出租车乘车概率预测方法的应用，还须给出如下的假设：

假设1 路网结构中各个道路的膨胀细化假设。

通过对现有路网数据中各个道路进行膨胀细化处理，合并路网结构中彼此间相近的道路，而原本归属于不同道路的出租车将视为在合并后的道路上。

假设2 出租车GPS数据与路网中各个道路关联性假设。

收集出租车GPS数据时，由于仪器等原因导致数据并非十分精确，从而对于未落在具体道路上的出租车GPS数据将采用最近原则，即出租车GPS数据点归属于距离其最近的道路。

假设3 概率预测方法的层次结构假设。

出租车乘车概率预测问题实质上是一个多属性决策问题，上一层中属性评价值的确定依赖于下一层子属性评价值的确定。这种依赖关系以规则库中规则的形式给出。如顶层的乘车概率评价值的确定，依赖于下一层的乘车地点和乘车时间评价值的确定。

3 乘车概率预测方法

出租车乘车概率预测方法的建模，主要是从路网数据和出租车GPS数据中挖掘出有效信息构建置信规则库，然后利用证据推理（evidential reasoning，ER）算法[9-11]合成询问点的乘车概率，其具体包含以下4个步骤。

步骤1 预处理矢量模型的路网数据；

步骤2 匹配出租车GPS数据与路网数据；

步骤3 构建出租车乘车概率预测的置信规则库；

步骤4 利用置信规则库推理询问点的出租车乘车概率。

上述步骤中路网数据与出租车GPS数据的数据量级较大，需要花费大量的时间，因此针对步骤1至步骤3的计算过程采用离线处理的方式。而对于为乘客提供当前的乘车概率，需要满足实时性的要求，因此针对步骤4的计算过程则采用在线处理的方式。

3.1 预处理路网数据

路网数据的预处理是指将路网数据由繁琐转化为简便的数据类型，主要包括地图分割、栅格化、膨胀与细化3个部分。

3.1.1 地图分割

在路网数据的处理中，其涵盖的区域范围可能过大，因此需要对路网进行区域划分，即地图分割。地图分割的核心思想是：首先根据数据中地理或路段的特性制定相应的划分规则；然后依此将地图分割成大小合理的分块。一种简单的地图分割方式是在综合权衡块数和块大小的情况下将路网分成若干等大小的矩形。例如对北京市的路网数据进行地图分割，由于北京市横跨的经度范围为115.375 0至117.500 0，纬度范围为39.426 7至41.083 3，地图分割时选取每块的边长为0.03，则总共可以将北京市的路网数据分割成3 976块。此外，目前可行的地图分割方法还有基于主干公路的地图分割算法等。在地图分割过程中，初始路网数据中道路的表示采用的是矢量点，而地图分割会造成块与块之间道路连接信息的丢失，因此需要特别处理相邻块的道路连接部分。假设存在 A 和 B 两个表示道路的矢量点，经地图分割后 A 和 B 分别处在不同的分块中，由于只有当 A 和 B 两个矢量点同时表示道路时才能确保道路信息完整，而地图分割会造成 A 点到边界及边界到 B 点的道路无法表示，需要利用斜率推算出道路的方向，以近似的方式还原道路信息。此外，地图分割还能有效地提升轨迹分析和时空数据挖掘的效率，即通过对地图进行区域划分能使每个块间的路网数据相互独立，因此当数据量级较大时，可采用现有的并行技术对各个分块进行并行处理。

3.1.2 栅格化

在地理信息系统里，通常使用两种模型表示地理信息，即矢量模型（vector-based model）和栅格模型（raster-based model）。矢量模型主要是依据笛卡尔坐标来确定点、线和多边形的位置，从而表示不同的地理数据对象；栅格模型则是首先将地理区域分成若干的小格子，利用地理对象所占据的各自的编号来存储地理对象的位置信息。现有的路网数据常用

矢量模型表示道路信息,但在地理区域分析中,矢量模型下的计算代价非常大,例如在矢量模型下的地图简化问题已经被证明是NP难问题[12],因此需要将其转换为更易于处理地理区域分析的栅格模型。对于模型的转换,可采用Bresenham算法[13]。该算法是计算机图形学领域使用最广泛的直线扫描转换方法,其核心思想是:通过各行、各列像素中心构造一组虚拟网格线,按直线从起点到终点的顺序计算直线各垂直网格线的交点,然后确定该列像素中于此交点最近的像素。下面以矢量模型的路网数据转换为栅格模型的路网数据为例介绍算法的基本流程。

Input:路网数据中的矢量点集 $N_1, N_2, \cdots, N_m$

Output:二维栅格模型

1. for N_i, N_{i+1} from N_1 to N_m
2. boolean $steep = \text{abs}(N_{i+1}.y - N_i.y) > \text{abs}(N_{i+1}.x - N_i.x)$;
3. if $steep ==$ true then
4. $swap\ (N_i.x, N_i.y)$;
5. $swap\ (N_{i+1}.x, N_{i+1}.y)$;
6. end
7. if $N_i.x > N_{i+1}.x$ then
8. $swap\ (N_i.x, N_{i+1}.x)$;
9. $swap\ (N_i.y, N_{i+1}.y)$;
10. end
11. int $deltax = N_{i+1}.x - N_i.x$, $deltay = \text{abs}(N_{i+1}.y - N_i.y)$;
12. $double\ error = 0$, $deltaerr = deltay/deltax$;
13. int $ystep = N_i.y < N_{i+1}.y\ ?1:-1$;
14. int $y = N_i.y$;
15. for x from $N_i.x$ to $N_{i+1}.x$
16. if $steep$ then $map[y][x] = 1$ else $plot[x][y] = 1$;
17. end
18. $error = error + deltaerr$;
19. if $error \geqslant 0.5$ then
20. $y = y + ystep$;
21. $error = error - 1.0$;
22. end
23. end
24. end

3.1.3 膨胀与细化

经栅格化处理后,路网数据可以看作一个仅有0-1的二值位图,其中1代表道路,0代表非道路。然而,此时栅格模型的地图中还存在诸如一些相近车道被视为不同道路等的细节,其会对后续处理造成不必要的时间与空间上的开销,因此需要通过膨胀和细化操作去掉路网数据中的细节。其中膨胀的作用是对每个点拓展出可能的连通点,使得位图上未连通但在实际上连通的部分相连。膨胀操作的基本原理可表示如下:设 A 是原二值位图像素点的集合,B 是结构元,其同样也是一个二值位图,则 A 相对 B 的膨胀可定义为:

$$A \oplus B = \bigcup_{b \in B} A_b \tag{1}$$

其中,$A_b = \{a + b | a \in A\}$,即将 A 整体平移向量 b 个单位。膨胀操作是具有交换性的,因此还可定义为:

$$A \oplus B = B \oplus A = \bigcup_{a \in A} B_a \tag{2}$$

事实上,对于任意像素 p,膨胀后 $p=1$,当且仅当平移到以 p 为原点的 B 与 A 的交集非空。而在膨胀过程中,通常采用3×3的元素全为1的方阵作为结构元。换言之,膨胀操作可简单地视为对每个值为1的像素填补其相邻8个位置的像素值为1。

膨胀操作后,虽然有效消除了路网数据中不必要的细节信息,但也使原路网中道路的宽度被过于夸张地放大,因此还需要对栅格模型的地图进行细化操作。细化的作用是将位图中数据膨胀出的无用部分消去,最小化位图搜索样本。现有可行的细化算法种类较多,但多数算法都需要通过一定次数的迭代来完成细化操作,算法的时间复杂度受迭代次数的影响。为兼顾处理路网数据的效率,本文采用文献[14]提出的快速并行细化算法来细化膨胀后的路网数据。以下对该算法进行简单介绍。

Input:如图1所示的八邻域点的二值位图

Output:细化后的二值位图

1. if $p_1 == 0$ then
2. return;
3. end
4. $N(p_1) \leftarrow$ 以 p_1 为中心的非零邻点的个数
5. $S(p_1) \leftarrow$ 在 $p_2, p_3, \cdots, p_9$ 中相邻点为0-1序列的个数;

6. if $2 \leqslant N(p_1) \leqslant 6$ and $S(p_1) == 1$ then
7. if $(p_2 \times p_4 \times p_6 == 0$ and $p_4 \times p_6 \times p_8 == 0)$
or $(p_2 \times p_4 \times p_8 == 0$ and $p_2 \times p_6 \times p_8 == 0)$ then
8. $p_1 = 0$;
9. end
10. end

p_9	p_2	p_3
p_8	p_1	p_4
p_7	p_6	p_5

Fig.1 Binary image of eight neighborhood points

图1 八邻域点的二值位图

3.2 匹配GPS数据

匹配GPS数据的目的是将出租车GPS数据分配到路网数据的各个分块中，并以合理的方式存储数据。在数据匹配过程中，为方便高效地访问GPS数据，首先将GPS数据依照路网数据中的分割方式分块存储，然后再将各块的GPS数据依照时间顺序分段存储。其中分段存储中，每段的时间间隔依实际情况而定，例如以3 min为界，则将在0时0分0秒至0时2分59秒间的数据划分为一段，0时3分0秒至0时5分59秒间的数据划分为一段。在时间间隔的选择上，当时间间隔越大，则每次访问特定GPS数据的所需时间越多；反之GPS数据的管理越复杂。此外，由于GPS车载接收机、交通矢量图等因素可能会引起误差，造成GPS数据并非准确地落在路网的道路上，还要将每条GPS数据匹配到路网的道路上。针对上述可能出现的问题，以下简要介绍如何将GPS数据匹配到路网上，并将GPS数据转换为用地图块号和块内坐标表示的数据类型。

Input：未预处理的出租车GPS数据集 $s_1, s_2, \cdots, s_m$

Output：由地图的分块编号 (*blockId*) 与块内坐标 (x, y) 组成的GPS数据

1. for s_i from s_1 to s_m
2. if s_i 不是有效的GPS数据 then
3. continue;
4. end
5. *blockId* ← 依据 s_i 的经纬度计算地图分块编号;
6. map_i ← 依据 *blockId* 获取 s_i 所在的分块地图;
7. (x_i, y_i) ← 依据 map_i 和 s_i 的经纬度计算分块地图中 s_i 的块内坐标;
8. if (x_i, y_i) 不属于路网道路上的坐标点 then
9. 搜索与 (x_i, y_i) 最近且位于路网道路上的坐标 (x'_i, y'_i);
10. $(x_i, y_i) \leftarrow (x'_i, y'_i)$;
11. end
12. end

假设现有北京市某一出租车GPS数据的经纬度为(116.214 996 3，39.730 419 2)，则可以依据北京市所在的经纬度范围及地图分块中每块的间隔求出地图分块编号：首先由GPS数据的纬度可确定其在沿经度方向的第 $\lceil (39.730\,419\,2 - 39.416\,667)/0.03 \rceil = 11$ 块，接着由GPS数据的经度可确定其在沿纬度方向的第 $\lceil (116.214\,996\,3 - 115.375\,000)/0.03 \rceil = 28$ 块，沿经度方向可将北京市地图分成 $\lceil (41.083\,333 - 39.416\,667)/0.03 \rceil = 56$ 块，因此最终可确定当GPS数据的经纬度为(116.214 996 3，39.730 419 2)时，其在地图分块中的第 $(11-1) \times 56 + 28 = 588$ 块。而对于块内坐标的转换，需要依据地图分块的映射函数。假设北京市路网中经度和纬度的映射函数分别是 $f(x) = \lfloor 100\,000 \times (x - 116.214\,996\,3) \rfloor \bmod 3\,000$，$f(y) = \lfloor 100\,000 \times (y - 39.416\,667) \rfloor \bmod 3\,000$，则转换后的块内坐标为(2 999，1 375)。

3.3 构建置信规则库

置信规则库模型[5]可用下面的四元组表示：

$$R = \langle U, A, D, F \rangle \tag{3}$$

其中，$U = \{U_i; i = 1, 2, \cdots, T\}$ 是规则的前件属性集合；每一个前件属性都有相应的候选值集合 $A = \{A_1, A_2, \cdots, A_T\}$，候选值集合中每个元素表示为 $A_i = \{A_{i,j}; j = 1, 2, \cdots, J_i\}$，$A_{i,j}$ 可以是定量的参考值，也可以是定性的评价等级，在规则库中，每条规则通常只包含前件属性的候选值集合中的一个元素，不同前件属性间候选值可以由逻辑关系“与”和“或”连接，其符号形式表示为 $\wedge$ 和 $\vee$；$D = \{D_n; n = 1, 2, \cdots, N\}$

是规则的后件属性，D_n 表示分布式的评价等级；F 是反映前件属性与后件属性之间关系的逻辑函数。在此基础上，置信规则的基本形式表示如下：

$$R_k: \text{if } A_1^k \wedge A_2^k \wedge \cdots \wedge A_{T_k}^k \text{ then } \{(D_1,\beta_{1,k}),(D_2,\beta_{2,k}),\cdots,(D_N,\beta_{N,k})\} \quad (4)$$

其中，$R_k(k=1,2,\cdots,L)$ 表示第 k 条规则；A_i^k 是第 i 个前件属性在第 k 条规则中的候选值，即 $A_i^k \in A_i$；T_k 是第 k 条规则中前件属性的数量；$\beta_{n,k}$ 是第 k 条规则中后件属性在评价等级 D_n 上的置信度，若 $\sum_{n=1}^{N}\beta_{n,k}<1$，称第 k 条规则包含的信息是完整的，否则称第 k 条规则包含的信息是不完整的。为提升规则表示不确定信息的能力，置信规则中还配有权重参数，其中 $\delta_{k,i}(i=1,2,\cdots,T_k)$ 是第 k 条规则中第 i 个前件属性的属性权重；θ_k 是第 k 条规则的规则权重。

对于出租车乘车概率预测的置信规则库的构建，首先需要从出租车GPS数据和路网数据中抽象出若干特征作为前件属性；然后依据前件属性的候选值及结合匹配后的路网数据和出租车GPS数据计算乘车概率，其中计算方式为在指定区域内统计出租车的载客数和空驶数，并以此计算乘车概率；最后利用上述信息创建带置信结构的置信规则。诸如以经纬度、时间作为前件属性，则有如下置信规则：当经纬度为(116.375 0,40.426 7)，时间为2012年11月1日14点05分00秒时，乘车概率为{(困难,0.2),(一般,0.5),(容易,0.3)}，且该条规则在规则库中的规则权重为0.7，各项前件属性的属性权重为0.2和0.6。

3.4 推理乘车概率

出租车乘车概率的实时计算需要依靠置信规则库推理方法，该方法的推理过程分成两部分，分别为计算激活权重和合成激活规则。对于激活权重的计算，首先根据输入值，假设置信规则库中第 i 个前件属性的输入值为 x_i 且其候选值集合为 $A_i=\{A_{i,j};j=1,2,\cdots,J_i\}$，则依据基于规则的信息转化技术[15]，可计算个体匹配度为：

$$\alpha_{i,j}=\begin{cases}\dfrac{A_{i,k+1}-x_i}{A_{i,k+1}-A_{i,k}}, j=k \text{ and } A_{i,k}\leqslant x_i\leqslant A_{i,k+1}\\ \dfrac{x_i-A_{i,k}}{A_{i,k+1}-A_{i,k}}, j=k+1 \text{ and } A_{i,k}\leqslant x_i\leqslant A_{i,k+1}\\ 0, j=1,2,\cdots,J_i \text{ and } j\neq k,k+1\end{cases} \quad (5)$$

当算得所有候选值的个体匹配度后，可用更直观的分布式框架表示个体匹配度：

$$S(x_i)=\{(A_{i,j},\alpha_{i,j});i=1,2,\cdots,T;j=1,2,\cdots,J_i\} \quad (6)$$

其中，$\alpha_{i,j}$ 是第 i 个前件属性中第 j 个候选值 $A_{i,j}$ 的个体匹配度；在概率预测的置信规则库中，假设前件属性时间的候选值包括14点00分00秒和14点15分00秒，当输入值为14点05分00秒时，候选值14点00分00秒对应的个体匹配度为0.666 7，而候选值14点15分00秒对应的个体匹配度则为0.333 3。

接着，结合权重参数可计算第 k 条规则的激活权重：

$$\omega_k=\frac{\theta_k\prod_{i=1}^{T_k}(\alpha_i^k)^{\bar{\delta}_{k,i}}}{\sum_{l=1}^{L}\left(\theta_l\prod_{i=1}^{T_k}(\alpha_i^l)^{\bar{\delta}_{l,i}}\right)}, \bar{\delta}_{k,i}=\frac{\delta_{k,i}}{\max\limits_{i=1,2,\cdots,T_k}\{\delta_{k,i}\}} \quad (7)$$

其中，α_i^k 是第 k 条规则中第 i 个前件属性候选值的个体匹配度。当激活权重 $\omega_k>0$ 时，表示第 k 条规则被激活。当求得置信规则库中所有激活规则的激活权重后，便可用证据推理算法中的解析公式[8]将所有激活规则一次合成，其各个评价等级上置信度的合成公式如下：

$$\beta_i=\frac{\prod_{k=1}^{L}\left(\omega_k\beta_{i,k}+1-\omega_k\sum_{i=1}^{N}\beta_{i,k}\right)-\prod_{k=1}^{L}\left(1-\omega_k\sum_{i=1}^{N}\beta_{i,k}\right)}{\sum_{i=1}^{N}\prod_{k=1}^{L}\left(\omega_k\beta_{i,k}+1-\omega_k\sum_{j=1}^{N}\beta_{j,k}\right)-(N-1)\prod_{k=1}^{L}\left(1-\omega_k\sum_{j=1}^{N}\beta_{j,k}\right)-\prod_{k=1}^{L}(1-\omega_k)} \quad (8)$$

当与前件属性相对应的输入值为 $x=\{x_1,x_2,\cdots,x_M\}$ 时，由上式可计算得到置信规则库的分布式输出：

$$f(x)=\{(D_i,\beta_i(x)),i=1,2,\cdots,N\} \quad (9)$$

为让置信规则库的输出更直观，假设后件属性中各个评价等级的等级效用值为 $\mu=\{\mu(D_1),\mu(D_2),\cdots,\mu(D_N)\}$，则可进一步推得置信规则库的数值型输出：

$$f_{\mu}(x)=\sum_{i=1}^{N}(\mu(D_i)\beta_i(x)) \tag{10}$$

4 示例

下面以北京市12 000辆出租车在2012年11月某天产生的所有GPS数据为例说明本文方法的应用。其中北京市路网数据获取地址为http://www.datatang.com/data/43855；GPS数据获取地址为http://www.datatang.com/data/44502。首先依据北京市路网数据，以每块的边长为0.03将北京市分割成3 976块分块地图；然后再将出租车GPS数据依据地图分块进行分块存储，并以3 min作为时间间隔对各个分块的GPS数据进行分段存储，进而提高处理数据的效率。在构建置信规则库中，选用经纬度 c_1 和时间 c_2 作为前件属性，并在每个分块中确定固定位置统计出租车的乘车概率，其中相邻的固定点相差0.4 km，因此统计乘车概率时以15 min为间隔依次统计在以固定点为中心0.4 km范围以内出租车的空车数和非空车数，并以此求得该点的乘车概率。对于乘车的难易程度划分成困难、一般、容易3个定性等级，其相应的定量值分别为0、0.5、1.0。通过对出租车GPS数据和路网数据离线处理后，最终可构建拥有6 167 232条规则的置信规则库。此外，假设置信规则库中规则权重相等，规则的前件属性的属性权重也相等，即 $\theta_k=1.0(k=1,2,\cdots,6\ 167\ 232)$，$\delta_{k1}=\delta_{k2}$。实时统计询问点概率及在线推理询问点概率的实验环境为：Intel® Core™ i5-4570CPU，4 GB RAM，64位Windows操作系统，Microsoft Visual C++ 6.0。在此基础上，以下将具体分析如何利用置信规则库推理方法（belife rule-base inference methodology using evidential reasoning，RIMER）及结合固定点的乘车概率推理任意询问点的乘车概率，并以实时统计询问点的概率和所用时间作为比较对象。

假设当前询问点的信息为：北京市辽金城博物馆，经纬度为(116.365 980，39.867 970)，询问时间为12:03，则由路网数据可知距离询问点相邻的已知点有(116.366 740，39.867 887)、(116.366 750，39.869 777)、(116.365 070，39.867 787)，已知点中包含的规则信息如表1所示。

利用表1中的置信规则库推理乘车概率的主要过程是：首先，路网道路上的询问点距离3个已知点的距离分别为200.116 1 m、209.702 5 m、230.073 6 m，则可计算前件属性的个体匹配 $S\{(116.365\ 980, 39.867\ 970)\}=\{(c_{1,1}, 0.343\ 633), (c_{1,2}, 0.336\ 142), (c_{1,3},$

Table 1 Part of belief rule-base

表1 部分置信规则库

规则编号	前件条件	后件结果
1	$c^1_{1,1}=(116.366\ 740, 39.867\ 887)\wedge c^1_{2,1}=11{:}45{:}00$	{(困难，0.017 544)，(一般，0.982 456)，(容易，0)}
2	$c^2_{1,1}=(116.366\ 740, 39.867\ 887)\wedge c^2_{2,2}=12{:}00{:}00$	{(困难，0)，(一般，0.990 476)，(容易，0.009 524)}
3	$c^3_{1,1}=(116.366\ 740, 39.867\ 887)\wedge c^3_{2,3}=12{:}15{:}00$	{(困难，0)，(一般，0.991 870)，(容易，0.008 130)}
4	$c^4_{1,1}=(116.366\ 740, 39.867\ 887)\wedge c^4_{2,4}=12{:}45{:}00$	{(困难，0.160 714)，(一般，0.839 285)，(容易，0)}
5	$c^5_{1,2}=(116.366\ 750, 39.869\ 777)\wedge c^5_{2,1}=11{:}45{:}00$	{(困难，0.046 358)，(一般，0.953 642)，(容易，0)}
6	$c^6_{1,2}=(116.366\ 750, 39.869\ 777)\wedge c^6_{2,2}=12{:}00{:}00$	{(困难，0.052 632)，(一般，0.947 368)，(容易，0)}
7	$c^7_{1,2}=(116.366\ 750, 39.869\ 777)\wedge c^7_{2,3}=12{:}15{:}00$	{(困难，0.020 690)，(一般，0.979 310)，(容易，0)}
8	$c^8_{1,2}=(116.366750, 39.869\ 777)\wedge c^8_{2,4}=12{:}45{:}00$	{(困难，0.152 318)，(一般，0.847 682)，(容易，0)}
9	$c^9_{1,3}=(116.365\ 070, 39.867\ 787)\wedge c^9_{2,1}=11{:}45{:}00$	{(困难，0.026 548)，(一般，0.973 452)，(容易，0)}
10	$c^{10}_{1,3}=(116.365\ 070, 39.867\ 787)\wedge c^{10}_{2,2}=12{:}00{:}00$	{(困难，0)，(一般，0.990 476)，(容易，0.009 524)}
11	$c^{11}_{1,3}=(116.365\ 070, 39.867\ 787)\wedge c^{11}_{2,3}=12{:}15{:}00$	{(困难，0)，(一般，0.992 870)，(容易，0.008 130)}
12	$c^{12}_{1,3}=(116.365\ 070, 39.867\ 787)\wedge c^{12}_{2,4}=12{:}45{:}00$	{(困难，0.171 172)，(一般，0.828 828)，(容易，0)}
…	…	…

Table 2 Comparison of 10 points for Beijing

表2 北京市10个地点的比较结果

编号	地点	经纬度	时间	统计概率	统计用时/s	推理概率	推理用时/s
1	田村山南路	(116.241 010, 39.925 657)	07:00	0.285 714	9.463	0.289 1	0.046
2	水仙西路	(116.697 200, 39.918 147)	09:48	0.352 941	8.237	0.277 7	0.031
3	玉林南路	(116.365 990, 39.868 807)	04:18	0.545 455	9.297	0.700 8	0.047
4	陶然花园酒店	(116.392 460, 39.879 367)	00:54	0.358 025	10.125	0.437 9	0.063
5	天安门东	(116.408 960, 39.915 347)	07:00	0.330 275	10.608	0.331 3	0.047
6	皇城根遗址公园	(116.411 110, 39.917 627)	05:48	0.548 780	9.953	0.570 7	0.047
7	中关村三桥	(116.330 630, 39.992 707)	04:18	0.384 615	8.705	0.394 1	0.047
8	西大望南路	(116.486 160, 39.867 727)	22:24	0.545 455	10.209	0.559 9	0.048
9	化工路	(116.510 360, 39.882 457)	07:00	0.577 778	4.868	0.564 6	0.047
10	金蝉西路	(116.510 820, 39.879 077)	22:24	0.470 588	9.906	0.480 6	0.031

0.320 225)}, $S\{12{:}03\} = \{(c_{1,1}, 0), (c_{1,2}, 0.8), (c_{1,3}, 0.2), (c_{1,4}, 0)\}$;接着,计算各个规则的激活权重 $\omega_1=0$, $\omega_2=0.274\,906$, $\omega_3=0.068\,727$, $\omega_4=0$, $\omega_5=0$, $\omega_6=0.268\,914$, $\omega_7=0.067\,228$, $\omega_8=0$, $\omega_9=0$, $\omega_{10}=0.256\,180$, $\omega_{11}=0.064\,045$, $\omega_{12}=0$,因此可确定第2、3、5、6、10、11条规则被激活;然后,根据式(8)计算得分布型乘车概率为{(困难,0.010 033),(一般,0.986 124),(容易,0.003 876)};最后,由式(10)计算得数值型乘车概率为0.496 9。此外,在本文方法的推理过程中,从给出询问点至最终得出概率所需的时间为0.072 s;相应的,通过对询问点实时统计乘车概率所需的时间为15.813 s,统计的乘车概率为0.551 282。由此可见,本文方法推理所得的结果与统计所得的乘车概率近乎相等,但所用的时间明显少于统计所需的时间。

为进一步验证本文方法的准确性和实时性,从北京路网中选取10个位置用于比较实时统计的概率和推理所得的概率,如表2所示。由表2可知,利用置信规则库推理的乘车概率与统计的乘车概率基本接近,其中编号3的玉林南路中误差最大,差值为0.155 345,而编号1的田村山南路的误差最小,差值为0.003 360。在耗时方面,由于本文方法包含离线处理,使用时间明显优于实时统计所需的时间。综上可得,本文方法能够较准确地推理出询问点的乘车概率,且相比实时统计的方式,更满足实时性的要求。

5 结束语

本文提出的基于置信规则库推理的出租车乘车概率在线预测方法是一种离线处理与在线处理相结合的方法。与常见的概率预测方法相比,本文方法能够通过构建置信规则库来实时推理得到路网任意位置上的乘车概率,且在数据处理过程中,其并非仅是一个简单的数值运算,而是在允许数据包含不确定信息的情形下自底向上推理乘车概率。在以2012年11月份北京出租车GPS数据为例的实验中,其结果充分说明了本文方法能够满足实时性和准确性的要求。由本文方法构建的置信规则库涉及的规则数众多,应用结构优化方法精简规则数将是下一步研究的重点。

References:

[1] Arel I, Liu C, Urbanik T, et al. Reinforcement learning-based multi-agent system for network traffic signal control[J]. IET Intelligent Transport Systems, 2010, 4(2): 128-135.

[2] Zhang Hesheng, Zhang Yi, Wen Huimin, et al. Estimation approaches of average link travel time using GPS data[J]. Journal of Jilin University: Engineering and Technology Edition, 2007, 37(3): 533-537.

[3] Yuan Jing, Zheng Yu, Zhang Liuhang, et al. T-Finder: a recommender system for finding passengers and vacant taxis[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(10): 2390-2403.

[4] Jiang Guiyan, Chang Ande, Li Qi, et al. Estimation models for average speed of traffic flow based on GPS data of taxi[J]. Journal of Southwest Jiaotong University, 2011, 46(4): 638-644.

[5] Yang Jianbo, Liu Jun, Wang Jin, et al. Belief rule-base inference methodology using the evidential reasoning approach—RIMER[J]. IEEE Transaction on Systems, Man, and Cybernetics: Part A Systems and Humans, 2006, 37(4): 569-585.

[6] Cheng Ben, Jiang Jiang, Tan Yuejin, et al. A novel approach for WSoS capability requirement satisfactory degree evaluation using evidential reasoning[J]. System Engineering Theory & Practice. 2011, 31(11): 2210-2216.

[7] Yang Longhao, Fu Yanggeng, Wu Yingjie. Structure learning approach of belief rule base for best decision structure[J]. Journal of Frontiers of Computer Science and Technology, 2014, 8(10): 1216-1230.

[8] Wu Weikun, Yang Longhao, Fu Yanggeng, et al. Parameters training for belief rule-base using the accelerating of gradient algorithm[J]. Journal of Frontiers of Computer Science and Technology, 2014, 8(8): 989-1001.

[9] Yang Jianbo, Xu Dongling. On the evidential reasoning algorithm for multiple attribute decision analysis under uncertainty[J]. IEEE Transactions on Systems, Man, and Cybernetics: Part A Systems and Humans, 2002, 32(3): 289-304.

[10] Wang Yingming, Yang Jianbo, Xu Dongling. Environmental impact assessment using the evidential reasoning approach[J]. European Journal of Operational Research, 2006, 174(3): 1885-1913.

[11] Fu Yanggeng, Yang Longhao, Wu Yingjie. Evidential reasoning approach for solving complex evaluation models[J]. Pattern Recognition and Artificial Intelligence, 2014, 27(4): 313-326.

[12] Estkowski R. No Steiner point subdivision simplification is NP-complete[C]//Proceedings of the 10th Canadian Conference Computational Geometry, 1998: 76-77.

[13] Bresenham J E. Algorithm for computer control of a digital potter[J]. IBM Systems Journal, 1965, 4(1): 25-30.

[14] Zhang T Y, Suen C Y. A fast parallel algorithm for thinning digital patterns[J]. Communications of the ACM, 1984, 27(3): 236-239.

[15] Yang Jianbo. Rule and utility based evidential reasoning approach for multi-attribute decision analysis under uncertainties[J]. European Journal of Operational Research, 2001, 131(1): 31-61.

附中文参考文献：

[2] 张和生, 张毅, 温惠敏, 等. 利用GPS数据估计路段的平均行程时间[J]. 吉林大学学报: 工学版, 2007, 37(3): 533-537.

[4] 姜桂艳, 常安德, 李琦, 等. 基于出租车GPS数据的路段平均速度估计模型[J]. 西南交通大学学报, 2011, 46(4): 638-644.

[6] 程贲, 姜江, 谭跃进, 等. 基于证据推理的武器装备体系能力需求满足度评估方法[J]. 系统工程理论与实践, 2011, 31(11): 2210-2216.

[7] 杨隆浩, 傅仰耿, 吴英杰. 面向最佳决策结构的置信规则库结构学习方法[J]. 计算机科学与探索, 2014, 8(10): 1216-1230.

[8] 吴伟昆, 杨隆浩, 傅仰耿, 等. 基于加速梯度求法的置信规则库参数训练方法[J]. 计算机科学与探索, 2014, 8(8): 989-1001.

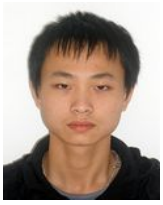
[11] 傅仰耿, 杨隆浩, 吴英杰. 面向复杂评价模型的证据推理方法[J]. 模式识别与人工智能, 2014, 27(4): 313-326.

YANG Longhao was born in 1990. He is a master candidate at College of Mathematics and Computer Science, Fuzhou University. His research interests include intelligent decision-making technology and belief rule-base inference, etc.

杨隆浩(1990—),男,福建南平人,福州大学数学与计算机科学学院硕士研究生,主要研究领域为智能决策技术,置信规则库推理等。

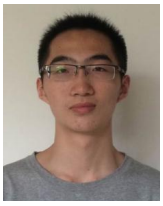
CAI Zhiling was born in 1992. She is a student at College of Mathematics and Computer Science, Fuzhou University. Her research interests include intelligent decision-making technology and data mining, etc.

蔡芷铃(1992—),女,福建晋江人,福州大学数学与计算机科学学院学生,主要研究领域为智能决策技术,数据挖掘等。



HUANG Zhixin was born in 1991. He is a student at College of Mathematics and Computer Science, Fuzhou University. His research interests include intelligent decision-making technology and data mining, etc.

黄志鑫(1991—),男,福建漳州人,福州大学数学与计算机科学学院学生,主要研究领域为智能决策技术,数据挖掘等。



HE Xing was born in 1992. He is a student at College of Mathematics and Computer Science, Fuzhou University. His research interests include intelligent decision-making technology and data mining, etc.

何星(1992—),男,福建福清人,福州大学数学与计算机科学学院学生,主要研究领域为智能决策技术,数据挖掘等。

FU Yanggeng was born in 1981. He received the Ph.D. degree from Fuzhou University in 2013. Now he is a lecturer at College of Mathematics and Computer Science, Fuzhou University, and the member of CCF. His research interests include multi-criteria decision-making under uncertainty, belief rule-base inference and mobile Internet applications, etc.

傅仰耿(1981—),男,福建泉州人,2013年于福州大学获得博士学位,现为福州大学数学与计算机科学学院讲师,CCF会员,主要研究领域为不确定多准则决策,置信规则库推理,移动互联网应用等。