AMUR:一种 RFID 数据不确定性的自适应度量算法

王永利[1],钱江波[2],孙淑荣[3],张功萱[1],刘冬梅[1]
(1. 南京理工大学计算机科学与技术学院,江苏南京 210094; 2. 宁波大学信息科学与工程学院,浙江宁波 315211;
3. 国电南京自动化股份有限公司技术管控部,江苏南京 211100)

摘 要: 为适应基于 RFID(无线射频识别)位置跟踪过程中传感数据的连续变化和需要实时处理的特征,本文提出一种度量 RFID 数据不确定性的自适应进化粒子滤波算法,根据 K-L 距离改变重采样粒子个数,并引入粒子群寻优方法 PSO 改变传统粒子滤波(SIRPF)的重采样效率,采用常规赋权聚集(CWA)定义适应度函数,以均衡先验密度与似然密度的重要性,在采样粒子空间探寻最优粒子,为概率数据库上的初始元组提供可靠的置信度度量.实验证明,与已有的算法相比,AMUR 算法能够有效地度量 RFID 数据中蕴含的不确定性,可进一步改善粒子退化现象和粒子贫化问题.

关键词: 无线射频识别; 物联网; 不确定性; 粒子滤波; 自适应; 粒子群优化
中图分类号: TP311.2 **文献标识码:** A **文章编号:** 0372-2112 (2011) 03-0579-06

AMUR: An Adaptive Measuring Algorithm of Underlying Uncertainty for RFID Data

WANG Yong-li[1], QIAN Jiang-bo[2], SUN Shu-rong[3], ZHANG Gong-xuan[1], LIU Dong-mei[1]
(1. *School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing, Jiangsu* 210094, *China*;
2. *Department of Information and Engineering, Ningbo University, Ningbo, Zhejiang* 315211, *China*;
3. *Department of Technique Management, GuoDian Nanjing Automation limited corporation. Nanjing, Jiangsu* 211100, *China*)

Abstract: To adapt the character of evolving over time and real-time of sensor data in location tracing service based on RFID, we present an adaptive evolving particle filtering algorithm-AMUR(an adaptive measuring algorithm of underlying uncertainty for RFID data). AMUR adaptively changes the number of samples on the basis of K-L distance, introduces an improved PSO (particle swarm optimization) method to enhance the efficiency of resampling phase of conventional particle filter(SIRPF). Meanwhile, to detect the most optimal samples among candidate sample set, AMUR defines a fitness function based on CWA(conventional weighted aggregation) for PSO which balances the importance between priori density and likelihood densitys. It provides a reliable measure of confidence for initial tuple in the probability RFID database. Experimental comparison of current algorithms shows, AMUR outpreforms current methods in terms of measurement of underlying uncertainties over RFID data, particle degradation and particle depletion.

Key words: RFID; Internet of things; uncertainty; particle filter; adaptive; particle swarm optimization

1 引言

对象位置追踪是基于 RFID 的物联网应用的关键问题,RFID 传感器数据反应了对象的位置信息,这些数据的精度受到传感器自身及周围环境中信号散射、碰撞的影响[1],因此不确定性是 RFID 传感器数据具有的最基本特征,有效地处理 RFID 数据的不确定性对于提高对象定位的精度至关重要.

在物流 RFID 系统的应用中,经常会涉及诸如"第 X 号包裹在哪里? 从何处来? 经过了哪些中转? 延迟交付的可能性有多大?"的查询,这是典型的针对数据不确定性和数据血统(起源)的查询[2].传统关系数据库系统不考虑数据的不确定性,为追溯某一数据的起源信息可以设计多种等价的查询计划并且得到相同结果,然而对于概率数据库,由于可能实例(possible instance)的存在,不同查询计划返回查询结果的概率值却可能不同.其根本原因是设计查询计划的过程中未考虑到数据的概率之间的相关性,导致重复计算[3].为解决此问题,首先要在传感数据首次流入处理系统时为它们附加上出现的概率.

收稿日期:2009-12-21;修回日期:2010-09-12
基金项目:国家自然科学基金项目(No.60803001,No.60803021,No.60850002);中国博士后科学基金特别资助项目(No.200902517);江苏省博士后基金(No.0801043B);南京市科技计划项目(No.2010 软资 02014)

对 RFID 数据的不确定性度量研究目前比较少,相关的领域主要针对科学数据库和确定数据库,未考虑以数据流处理为特征的 RFID 数据不确定性进化特点,这使得现有的概率估计方法很难直接应用到 RFID 数据管理.与本文相近的工作是 ULDB,首次结合不确定性与血统的研究,文献[2]讨论了利用血统模拟数据的不确定性与查询结果,文献[3]研究怎样提高不确定数据上的处理效率,为了确定查询结果的置信度,采用元组上的全局布尔公式(记为血统)计算元组的概率.然而已有的工作都是假设原始数据概率已知(即预设好了不确定性),ULDB 未提及原始数据概率的度量,尤其是多目标传感器数据概率的度量问题.

在现代目标跟踪领域,由于实际问题的复杂性,所面对的更多的是非线性非高斯问题[4].Hue 等[5]把粒子滤波(PF)推广到多目标跟踪和数据关联.与本文工作近似的方法是 Dieter Fox 提出的 KLD-sampling 算法[6],根据最大似然估计与真实分布之间的 K-L 距离动态调整重采样粒子的个数,也可在一定程度上解决退化问题,但易导致样本贫化,难以适应状态变化的情况.在 RFID 应用中,Christopher 等[7]首先提出利用粒子滤波算法来为 RFID 数据添加概率维来表示 RFID 数据的不确定性,然而其方法中粒子滤波的重采样个数是固定的,无法适应状态空间随时间演化的特点,定位精度偏低.方正提出 PSOPF 算法[8]对采样过程利用粒子群优化算法进行优化,使粒子集向后验概率密度分布取值较大的区域运动,从而克服了粒子贫乏问题,然而其固定的粒子数目无法体现数据进化的特点,同时其粒子群的适应度函数没有考虑到偏重先验还是偏重似然的优化均衡问题.

本文针对对象位置追踪应用需求,基于 K-L 距离捕获多标签 RFID 数据的进化情况自适应先验样本的采样个数,并在此基础上提出一种优化重采样粒子质量的方法——AMUR(Adaptive Measuring Algorithm of Underlying Uncertainty for RFID Data),从不确定性本源上提供原始数据中蕴含的不确定性度量.AMUR 方法的贡献有二:(1)根据 RFID 数据进化情况自适应调整粒子数目,改善粒子退化现象;(2)采用改进的粒子群算法 PSO 优化重采样性能,改善粒子贫化问题.

2 模型与定义

2.1 不确定 RFID 数据的数据模型

在大多数应用中,数据的不确定性可细分为元组级不确定性和属性级不确定性.元组级不确定性描述元组的存在与否,较为通用.属性级不确定性并不涉及整个元组的不确定性,而是以概率密度函数或统计参数(例如方差等)来描述特定属性的不确定性.

定义 1 可能世界实例:各元组的任一合法组合均构成一个可能世界实例(instance),实例的概率值可以通过相关元组的概率计算得到,元组之间可能独立也可能存在依赖关系,存在元组代表实际的 possible-tuple 组合.

定义 2 RFID 数据流集 ψ:由多条数据流 S 组成,滑动窗口中的 S 部分相当于关系数据模型中的表,元组 $t \in S$,由多个属性值组成.

由于数据流潜在的无限性,不可能将所有的数据全部存储后才计算,因此必须确定合适的数据流窗口(滑动窗口),大小适中的滑动窗口即保留了窗口中的采样样本的统计完备性,又降低了数据流的存储代价.令 p_i^{avg} 表示在一个间隙中标签 i 的观测概率,若在平滑窗口中间隙的个数 n_i 满足不等式 $n_i \geqslant \left| \frac{\ln(\frac{1}{\rho})}{p_i^{\text{avg}}} \right|$,则可以保证在窗口 n_i 中以大于 $1-\rho$ 的概率读取标签 i[9].

2.2 粒子滤波

粒子滤波(Particle Filter)的思想基于蒙特卡洛方法(Monte Carlo methods),它利用粒子集来表示概率,可以用在任何形式的状态空间模型上.其核心思想是通过从后验概率中抽取的随机状态粒子来表达其分布,是序列蒙特卡罗滤波方法(SMC,Sequential Monte Carlo)的一种,它用一个含权的点集 $\{(x^i, w^j), i = 1,2,\cdots,N\}$ 近似系统分布概率.尽其概率分布只是真实分布的一种近似,但由于非参数化的特点,它摆脱了解决非线性滤波问题时随机量必须满足高斯分布的制约,能表达比高斯模型更广泛的分布,对变量参数的非线性特性有更强的建模能力.因此,粒子滤波能够比较精确地表达基于观测量和控制量的后验概率分布,适用于实际的 RFID 应用.根据 RFID 目标跟踪的应用背景,考虑如下带有倍增噪声的非线性模式[10].

$$x_t = f(x_{t-1})(1 + v_{t-1}) \tag{1}$$

$$y_t = h(x_t)(1 + u_t) \tag{2}$$

其中,$f(x_t) = 0.5x_t + \frac{25x_t}{1+x_t^2} + 8\cos(1.2t)$,$h(x_t) = \frac{x_t^2}{20}$,$v_t$ 和 u_t 是均值为零方差为 Q 与 R 的白噪声.

3 算法理论基础

3.1 自适应计算采样粒子个数

为提高粒子滤波的性能,适应采样粒子数目随时间动态的变化,本文提出一种自适应计算采样粒子个数的新方法:在 Kullback-Leibler(K-L)距离的基础上描述采样的极大似然估计(MLE)与真实后验概率密度的差异,根据这种差异感知状态的变化,在给定近似误差边界ϵ的限定下动态调节样本个数.K-L距离是两个概率

分布 p 和 q 之间的差异:

$$K(p,q)=\sum_x p(x)\log\frac{p(x)}{q(x)} \tag{3}$$

采用非参数极大似然估计方法确定采样个数的误差边界,由于真实后验可表示为由多个桶集合组成离散常数分布,允许应用卡方渐近收敛统计似然比率,因此本文借助 K-L divergence 测度来确定粒子个数 N 的边界:

$$N>\frac{1}{2\varepsilon}\chi^2_{k-1,1-\delta} \tag{4}$$

其中 ε 为 KL-divergence 给定的误差上界,$(1-\delta)$是自由度为 $k-1$ 卡方分布的分位数,N 表示支持的桶个数.

KLD[6]采用经验分布作为确定边界的依据,即假设样本来自于真实分布,这并不适合粒子滤波的情况.粒子滤波中样本来自于估价权值的重要性函数.重要性函数与真实分布之间匹配的质量是决定粒子精度的重要因素之一,由 KLD 采样确定的边界仅仅使用有关真实后验的信息,但是忽略了真实和估计分布之间的不匹配.

为解决 KLD 采样问题,我们需要量化在样本估计过程中发生退化的程度(样本来自于重要性函数),目标是从重要性函数和真实密度中发现等价样本的个数,能够从两者中捕捉相等数量的信息. MonteCarlo (MC)积分[11]提出相关数字效率(RNE)的概念,给出了量化采样(从重要性函数得出)影响的指标,RNE 思想是比较解的相对精度,样本既来自真实密度又来自预测密度,根据积分估计子的方差度量精度.

定理 1　当样本不是来自于真实分布而是来自于重要性函数时,K-L divergence 测度下需要采样的粒子个数 N,满足边界条件 $N_{IS}>\frac{\sigma^2_{IS}}{Var_p(x)}\frac{1}{2\varepsilon}\chi^2_{k-1,1-\delta}$,并能够增量计算.

证明:采用 MC 积分估计状态的均值($E_{\mathrm{MC}}(x)$),估计子的方差参照文献[12]给定:

$$Var[E^N_{\mathrm{MC}}(x)]=\frac{Var_p(x)}{N} \tag{5}$$

其中 N 表示来自于真实分布 $p(x)$的样本个数,下标 p 表示用于目标分布涉及的方差.设样本来自于某重要函数 $q(x)$,估计子的方差相对于重要采样(IS)的方差,由下式指定:

$$Var[E^N_{IS}(x)]=E_q((x-E_p(x))^2w(x)^2)/N_{IS}=\sigma^2_{IS}/N_{IS} \tag{6}$$

其中 $w(x)$对应 $p(x)/q(x)$,即重要采样(IS)的权值,N_{IS}为样本的个数,样本来自重要函数.

为获得相似级别的精度,两个估计子的方差应该相等,这样可以发现量化真实密度与估计密度的样本间的平衡关系:

$$N=\frac{N_{IS}Var_p(x)}{\sigma^2_{IS}} \tag{7}$$

将式(6)代入式(4),当样本不是来自于真实分布,而是来自于重要性函数时,可以纠正 KLD-采样给定的边界.

$$N_{IS}>\frac{\sigma^2_{IS}}{Var_p(x)}\frac{1}{2\varepsilon}\chi^2_{k-1,1-\delta} \tag{8}$$

使用 MC 积分,$Var_p(x)$与 σ^2_{IS}可以用下式估计:

$$Var_p(x)=E_p(x^2)-E_p(x)^2\approx\frac{\sum_{i=1}^N x_i^2w_i}{\sum_{i=1}^N w_i}-E_p(x)^2 \text{ 且}$$

$$\sigma^2_{IS}\approx\frac{\sum_{i=1}^N x_i^2w_i}{\sum_{i=1}^N w_i}-\frac{2\sum_{i=1}^N x_iw_i^2E_p(x)}{\sum_{i=1}^N w_i}+\frac{\sum_{i=1}^N w_i^2E_p(x)^2}{\sum_{i=1}^N w_i} \tag{9}$$

其中 $E_P(x)=\frac{\sum_{i=1}^n x_iw_i}{\sum_{i=1}^n w_i}$.等式(9)显示了采用合适的聚集,可以保持以 $O(N)$的复杂度增量计算边界.证毕.

3.2　优化重采样粒子

当状态空间的似然函数比较窄(即测量值比较精确)或紧密似然函数落在低先验密度区域(即先验函数的尾部)的时候,PF 会产生粒子贫化的问题,很有可能在预估结果中失去重要粒子.为解决粒子贫化问题,改变 PF 算法重采样阶段方式,在先验样本生成之后重采样之前引入一种改进的粒子群优化方法,将这些先验样本移至似然比较重要的状态空间区域,提高粒子的相异度.

受最大最小策略的适应值启发,将此问题描述为两个对象竞争的多目标问题.记第一个目标函数为 F_1,代表高似然区域的最大值,记第二个目标函数为 F_2,代表高先验区域的最大值.考虑到满足 PSO 算法的直接适用性、简洁性、及跟踪多目标问题的能力,采用常规赋权聚集(CWA)[13]方法合并目标,CWA 方法允许用户修改权值,有利于需要特别关注系统模式(先验)或观察模式(似然).假设已知两个目标函数,用 CWA 方法将两个目标合并为一个目标,得到如下最大化问题:

$$\max_{x\in S}F(x)=\alpha_1F_1(x)+\alpha_2F_2(x) \tag{10}$$

其中 α_1 和 α_2 为非负权值,且 $\alpha_1+\alpha_2=1$,S 为先验样本状态(搜索)空间.把生成的先验样本 $x_t^{(i)*}$,$i=1,2,\cdots,N$ 看作 PSO 的初始群,移动粒子目标是最大化目标函数 $F(x)$,借助 PSO 算法搜索得到由最佳位置组成的结果样本.对目前处于最佳位置的样本进行重采样,

重采样的粒子的个数根据分布的变化调整.

目标函数 F_1 和 F_2 与系统的似然和先验密切相关,本文的研究背景是 RFID 区域位置定位,用在等式(10)中的两个目标函数如下:

$$F_1(x_t^{(i),\mathrm{PSO}}) = \left|\frac{1}{h(x_t^{(i)^*})}\right| \frac{1}{\sqrt{2\pi R}}\exp\left\{-\frac{1}{2R}\left[\frac{y_t - h(x_t^{(i)^*})}{h(x_t^{(i)^*})}\right]^2\right\}$$

$$F_2(x_t^{(i),\mathrm{PSO}}) = \frac{1}{\sqrt{2\pi Q}}\exp\left\{-\frac{1}{2Q}\left[\frac{x_t^{(i),\mathrm{PSO}} - x_t^{(i)*}}{x_t^{(i)*}}\right]^2\right\}$$

其中 $x_t^{(i),\mathrm{PSO}}$ 是 PSO 中第 i 个群粒子,$x_t^{(i)*}$ 是 $x_t^{(i),\mathrm{PSO}}$ 初始化处所在的相应的采样点,y_t 是观测值,索引 t 代表粒子滤波系统的迭代节拍.

4 AMUR 算法描述

以 3.2 节构建的粒子群优化适应度函数为基础,在重采样之前对粒子优化的算法如图 1:

Procedure OptimizationParticlesSelection
Input: 给定群体规模N, 先验样本S_t, 测度噪声协方差 R_t
Output: 筛选之后的先验样本$s_t^{(i)*}$, i=1,2,…,N
(1)初始化粒子群:每个粒子对应的$P_{\mathrm{pbest}}=s_t^{(i)*}$, $v_t^{(i)}$=0
(2) **DO**
(3) **FOR** each particle i = 1, …,N
(4) fitness = $\max_{x\in S} F(x)=\alpha_1F_1(x)+\alpha_2F_2(x)$
(5) **IF** fitness > P_{pbest} THEN
(6) $v_t^{(i)}$=|Randn()| ($P_{\mathrm{pbest}}-s_t^{(i)*}$) + |Randn() |($P_{\mathrm{gbest}}-s_t^{(i)*}$)
(7) $s_{t+1}^{(i)*}=s_t^{(i)*}+v_t^{(i)}$
(8) Update P_{pbest}, P_{gbest}
(9) **ENDIF**
(10) **END FOR**
(11) **WHILE** termination condition not met
(12) **RETURN** $s_t^{(i)*}$, i=1,2,…,N

图1 重采样前粒子优化算法伪代码

通过以上的优化过程,使得粒子集在权重值更新前更加趋向于高似然区域,从而解决了粒子贫乏问题.同时,优化过程使得远离真实状态的粒子趋向于真实状态出现概率较大的区域,提高了每个粒子的作用效果.

为满足 RFID 数据不确定性的在线度量,为粒子滤波提供实时性支持,设计 ComputingWindowSize(p^{avg}, ρ)过程,为 AMUR 自适应进化粒子滤波算法确定合适的滑动窗口大小.其输入为标签平均概率与置信度概率,输出为合适的窗口尺寸 W.

根据定理 1,由粒子滤波更新的前两个步骤估计预测分布 $p(x_t \mid x_{t-1}, v_{t-1})\,Bel(x_{t-1})$,为在采样期间统计计算桶的个数,对于每个生成的样本,判断其是否落入了一个空箱来估计 k,只要样本的个数超过式(8)指定的阈值时就可以停止采样.自适应进化粒子滤波算法 AMUR 的描述如图 2.

可从标准统计表中获得典型 δ 的 $z_{1-\delta}$ 值,在给定 ε 和 δ 后,只需确定抽样中状态数目 n,就可得到当前所需的样本数 N,这样就可保证在给定的概率为 $1-\delta$ 下最大似然估计和真实分布间的 K-L 距离小于 ε.

AMUR 粒子滤波算法使用固定桶记录支持桶的个数 k,以及记录适应度函数的表.有序的适应度函数表采用折半查找等快速方法,给定桶尺寸后桶的最大个数 k 是有限的,因此 AMUR 采样过程一定会终止.

复杂度分析:在每个 t 时刻,step2 中语句(3)计算时间窗过程的时间复杂度为 $O(1)$;根据定理 2,语句(5)~(16)的时间复杂度为 $O(N)$,step3 中语句(17)优化过程采用查表法实现适应度函数的计算,时间复杂度为 $O(N\log_2 N)$,语句(18)到(20)为 $O(N)$;step4 语句(21)是重采样步骤产生新样本$\{x_k^{(i)}\}_{i=1}^n$,服从离散概率分布 $p_r(x_k^{(i)} = x_k^j) = w_k^j$,因而重新获得的样本可以认为是符合上式分布的一系列独立同分布的样本,其权重是一样的,均设为 $1/n$,这样只需要通过复杂度为 $O(N)$ 的操作就可以完成重采样过程.综上,AMUR 算法一个时钟节拍的时间复杂度为 $O(N\log_2 N)$,可借助算法中各种参数的调节,支持对不确定性估计的精度与计算复杂度的折中,符合在线计算的需要.

Algorithm AMURParticles
Input: $S_{t-1}=\{<x_{t-1}^{(i)}, w_{t-1}^{(i)}> \mid i=1,...,n\}$, 表示带有权值的粒子集,控制测度 v_{t-1}, 观测值 y_t, 阈值 ε、δ, 桶的大小 Δ, 标签平均读取率 p^{avg}, 概率 ρ
Output: S_t
Step 1:
(1)$S_{t-1}:=\phi, n:=0, k:=0, \alpha:=0$ //初始化
(2)**BEGIN**:
Step 2:
(3) W:=ComputingWindowSize(p^{avg}, ρ)
(4) 对于落在W中的粒子
(5) **DO**
(6) 从权值给定的S_{t-1}的离散分布中采样索引$j(n)$ //生成样本
(7) 用$x_{t-1}^{(j(n))}$与v_{t-1}从 $p(x_t \mid x_{t-1}, v_{t-1})Bel(x_{t-1})$ 采样 $x_t^{(n)}$
(8) $w_t^{(n)} := p(z_t \mid x_t^{(n)})$; //计算重要性权值
(9) $r := r + w_t^{(n)}$; //更新规范化因子
(10) $S_t := S_t \cup \{<x_t^{(n)}, w_t^{(n)}>\}$ //向样本集中插入样本
(11) **IF** ($x_t^{(n)}$落入空桶b) **THEN** //更新带有支持度的桶个数
(12) k:=k+1
(13) b:=non-empty
(14) n:=n+1 //更新生成样本的个数
(15) **ENDIF**
(16) **WHILE** ($n < \frac{\sigma_{ES}^2}{Var_p(x)}\frac{1}{2\varepsilon}\chi^2_{k-1,1-\delta}$) //直到抵达K-L边界
Step 3:
(17) S_t:=OptimizationParticlesSelection(N, S_t, R_t) // 粒子群优化,选择最佳样本
(18) **FOR** (i:=1,…,n) **DO** //归一化重要性权值
(19) $w_t^{(i)} := w_t^{(i)} / \varepsilon$
(20) **ENDDO**
Step 4:
(21) $\left\{x_t^{(i)}, \frac{1}{n}\right\}_{i=1}^n := \left\{x_t^{(i)}, w_t^{(i)}\right\}_{i=1}^n$//对处于最佳位置的样本进行重采样
(22) **IF** (t is not terminate) **GOTO BEGIN**

图2 自适应进化粒子滤波算法伪代码

5 实验评价

本文提出 AMUR 算法的应用场合为无环境电磁干扰的办公室环境下人员、物品的跟踪.实验中将办公楼层划分为若干个 $5M\times5M$ 的平面网格,视反应位置信息的 RFID 数据样本为粒子,布置 18 个阅读器,让携带标签的志愿者在这 18 个阅读器的识别范围内做随机匀速运动,阅读器向上位机提供的阅读器号表示物体的

位置信息,即 18 个阅读器分别提供整数 1 到 18 作为粒子滤波算法的输入.实验基于 Windows 2003 平台,上位机配置为 CPU:Intel CoreTM 2 Duo (2.9GHz)/主存:4GB.

采用等式(1)、(2)定义的模型测试本文提出的自适应优化粒子算法 AMUR, Q, $R \in \{0.01, 0.05, 0.1, 0.3, 0.5\}$,对应 1%、5%、10%、30%和 50%的扰动,代表不同的系统噪声和测试噪声级别,系统或测度值.滑动窗口采用 60 个时间步骤,$t = 1, 2, \cdots, 60$,每个时刻执行的粒子滤波过程中,在生成样本之前自适应计算粒子个数,生成样本之后触发粒子群优化.PSO 参数为 $\chi = 0.729$, $c_1 = c_2 = 2.05$,目标函数由等式(5)定义,优化阶段中采用三种不同的级别 α_1(α_2 简单的等于 $1 - \alpha_1$),0.2(先验重要一些),0.5(重要程度相等),0.8(似然重要一些).

设计带有时间戳的数据采集方法,允许实时评价系统状态,比较 AMUR 方法与 KLD[6]方法、PSOPF[8]方法、固定样本集尺寸的常规粒子(SIRPF)滤波方法,为四种方法设置可以比较的参数,对于 SIRPF 方法改变样本的数量,对于 PSOPF 方法改变阈值(阈值用来确定样本的个数),KLD 采用方法中改变 ε(即 K-L 距离的边界).实验 1~3 中 AMUR 方法的适应度参数 α_1 等于 2.0,实验 2 中 ρ 的值为 0.05,δ 的值为 0.99.

实验 1　真实后验的近似性能测试

实验 1 评价不同方法近似真实后验密度的精度.因无法获知后验的基础信息,用参照样本集与不同方法生成的样本集比较,用固定 10000 个样本的粒子滤波生成的参照集(远多于位置估计实际用的数目).每次迭代后,比较样本集与相应参照集之间的 K-L 距离,在两个集合上采用柱状图表示差异.

实验忽略了时间戳,算法用到的是需要处理数据的时间,图 3 描绘了不同算法的平均 K-L 距离(95%置信间隔)及平均样本数量,图中省略了超过 1.0 的大误差率,每一数据点表示标签(由人携带)在不同起始位置开始移动的 18 个全局位置的平均值,每次运行在不同的点时与 150 个样本集进行比较,正如预料中的一样,用的样本越多,近似效果越好.曲线图同时显示了本文提出方法的优异性能:常规粒子(SIRPF)滤波在汇聚到收敛 K-L 距离 0.3 之下时用到了 80000 个样本,本文提出的 AMUR 方法收敛到相同的水平(distance of K-L < = 0.3)仅平均需要 2000 个样本,KLD-Sampling 方法大约需要 10000 个样本,实验表明本文提出的方法尽管建立在几种近似之上,但仍然能跟踪真实的后验分布,适用于样本集较少的情况.

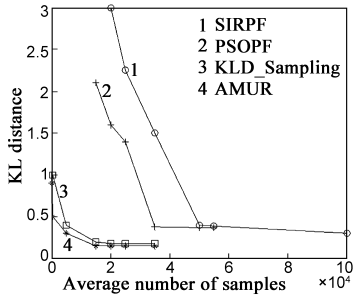

图3　不同平均样本数条件下KL距离的对比

实验 2　实时性能测试

因为确定样本个数及粒子群优化样本产生额外计算代价,实时条件下估计算法的性能至关重要.为测试真实条件下本方法的性能,执行数据流滑动窗口模型下的全局定位实验.记录 RFID 位置数据流集中的时间戳,改变 ε 边界的方式以获得不同的样本平均个数,边界 ε 上下界 0.35 和 0.010,对应于样本个数的最大值与最小值.

为评定不同算法的性能,每次迭代后确定估计的标签位置与相应于参照位置之间的距离.结果如图 3,其中 4 条曲线显示出来的 U 字形很好地诠释了样本数量选择代价与在实时约束下的平衡.样本选择不足导致对蕴含后验的不好的近似,经常无法确定标签的真实位置,另一方面,如果选择了过多的样本,算法的每次更新将持续几秒,从而不得不丢弃一些有价值的传感数据,这将导致位置估计的精度下降.图 3 显示即使是在实时情况下,AMUR 采样方法仍然能获得较好地性能提升,比固定大小尺寸与基于近似的采样方法性能好,略逊于 KLD 方法,因为度量先验与似然量级的开销.最小平均位置误差为 45cm,而另外三种方法分别为 60cm、80cm 与 120cm(固定),此结果的产生原因是,我们的方法考虑到了重要性函数与真实分布之间匹配的质量,可以确定最佳的折衷,而 KLD 采样确定边界仅仅使用有关真实后验的信息,在早期定位阶段用更多的样本,在位置追踪阶段用较少的样本.

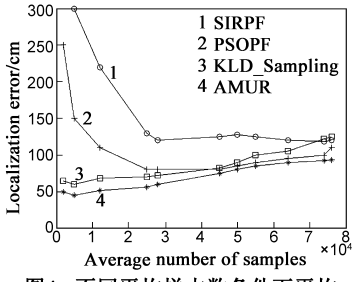

图4　不同平均样本数条件下平均定位误差对比

图 3 与图 4 中,x 轴表示四种不同的方法在可比较参数定义下的平均样本集尺寸,图 3 的 y 轴代表参照密度与不同方法生成样本集之间的 K-L 距离.图 4 的 y 轴代表平均位置误差,由测量估计位置与参考位置间距离之差计算得出,图 4 中呈现的 U 型是由于在实时条件下的原因,样本数量的增长导致了更高的更新次数,因此导致传感数据缺失.

实验 3　收敛性能测试

比较固定样本集尺寸的常规粒子滤波(SIRPF)与 AMUR 滤波的收敛性能,当采用相同粒子数(6000 个粒子)时,AMUR 粒子滤波方法很快地收敛到真实位置附近,而常规粒子滤波方法却无法在较短时间内收敛,甚至最终发散.同时测试另一组实验,比较随着粒子数下降情况下两种方法的性能,实验结果如表 1.可以看出,

当粒子数低于一定数目时,常规粒子滤波无法收敛到真实位置,而 AMUR 滤波却可利用较少的粒子收敛到真实位置附近.

本实验中,由于测试者的初始位置已知,AMUR 只需 50 个左右的粒子就可以实现精确的位置跟踪,而常规粒子滤波方法 SIRPF 大约需要 100 个以上,对于全局定位问题(50×20m 大小的环境),改进算法 AMUR 只需大约 500 个粒子就可以实现.

表 1 收敛效率对照表

样本大小	SIRPF/s	PSOPF/s	KLD/s	AMUR/s
6000	350	40	150	30
2500	发散	20	70	9
1250	发散	12	30	2
800	发散	5	发散	1
50	发散	发散	发散	0.1

6 结论

不确定性数据处理是国际上数据库领域近年来才兴起的热点研究方向,度量各种传感器采集的原始数据的不确定性是其中的关键问题.本文提出动态适应粒子滤波样本集大小的统计方法,用进化思想处理粒子退化,用粒子群算法提升重采样粒子的质量,解决了原 KLD 算法单纯依靠 K-L 距离确定粒子个数,引起的重采样中粒子有效性与多样性矛盾(即粒子贫乏)的问题.RFID 标签定位测试实验证明,本文提出的方法与已有方法相比可获得更好的位置估计精度,并且仅使用 SIRPF 方法 5%的样本量,KLDPF 方法 10%的样本量,获得了良好的性能提升,为概率数据库的近似解答、不确定信息血统追踪等上层操作提供了精确的原始不确定性度量,可广泛适用于物联网应用.下一步的工作是探讨其他状态方程和多维数据的不确定度量方法,以及追溯数据加工过程中不确定性的传播.

参考文献

[1] Derakhshan R, Orlowska M E, Li Xue. RFID data management: Challenges and opportunities[A]. Proceeding of IEEE International Conference on RFID[C]. Dallas: IEEE Computer Society, 2007. 175 – 182.

[2] Benjelloun O, Sarma A, Halevy A, Widom J. Uldbs: Databases with uncertainty and lineage[A]. Proceeding of the 32th International Conferance on Very Large Data Base (VLDB06)[C]. Seoul: VLDB Endowment, 2006. 953 – 964.

[3] Sarma A D, Theobald M, Widom J. Exploiting lineage for confidence computation in uncertain and probabilistic databases [A]. Proceedings of the 24th IEEE International Conference on Data Engineering[C]. Washington, DC: IEEE Computer Society Cancun, 2008. 1023 – 1032.

[4] Coates M. Distributed particle filters for sensor networks[A]. Proceedings of the 3rd International Conference on Information Processing in Sensor Networks (IPSN04)[C]. New York, NY: ACM, 2004. 99 – 107.

[5] Hue C, Cadre J, Perez P. Sequential Monte Carlo methods for multiple target tracking and data fusion [J]. IEEE Trans on Signal Processing, 2002, 50(2): 309 – 325.

[6] FOX D. Adapting the sample size in particle filters through KLD-Sampling[J]. The International Journal of Robotics Research, 2003, 22(12): 985 – 1003.

[7] Christopher Ré, Letchner J, Balazinksa M, Suciu D. Event queries on correlated probabilistic streams[A]. Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data[C]. New York, NY: ACM, 2008. 715 – 728.

[8] 方正,佟国峰,徐心和. 粒子群优化粒子滤波方法[J]. 控制与决策. 2007, 22(3): 273 – 277.
Fang Zheng, Tong Guo-Feng, Xu Xinhe. Particle swarm optimized particle filter[J]. Control and Decision. 2007, 22(3): 273 – 277. (in Chinese)

[9] Shawn R J, Minos G, Michael J F, Adaptive cleaning for RFID data streams[A]. Proceedings of the 32nd International Conference on Very Large Data Bases(VLDB06) [C]. Seoul: VLDB Endowment, 2006. 167 – 174.

[10] Gordon N J, Salmond D J, Smith A F M. Novel approach to nonlinear/non-gaussian bayesian state estimation[J]. IEE Proceedings F In Radar and Signal Processing, 2002, 140(2): 107 – 113.

[11] Geweke J. Bayesian inference in econometric models using Monte Carlo integration[J]. Journal of Econometrica, 1989, 57(6): 1317 – 1339.

[12] Doucet A, Freitas N D, Gordon N. Sequential Monte Carlo Methods in Practice[M]. New York, NY: Springer, 2001. 3 – 14.

[13] Parsopoulos K E, Vrahatis M N. Particle swarm optimization method in multiobjective problems[A]. Proceedings of the 2002 ACM Symposium on Applied Computing [C]. New York, NY: ACM, 2002. 603 – 607.

作者简介

王永利 男,1974 出生,副教授,博士,主要研究方向:传感数据处理,不确定数据数据处理,模式识别等.
E-mail: yongliwang@mail.njust.edu.cn

钱江波 男,1974 出生,男,副教授,博士,主要研究方向:数据库和数据流管理技术,数据挖掘,逻辑电路设计等.
E-mail: qianjiangbo@nbu.edu.cn

孙淑荣 女,1974 出生,工程师,硕士,主要研究方向:电力线通信技术,光通信,智能电网等.
E-mail: sunsr@sac-china.com