一种改进的动态 k-均值聚类算法①

胡 伟

(山西财经大学 实验教学中心, 太原 030006)

摘 要: 针对经典 k-均值聚类方法只能处理静态数据聚类的问题, 本文提出一种能够处理动态数据的改进动态 k-均值聚类算法, 称为 Dynamical K-means 算法. 该方法在经典 k-均值方法的基础上, 通过对动态变化的数据集中新加入样本进行分析和处理, 根据聚类目标函数改变的实际情况选择最相似的类别进行局部更新或进行全局经典 k-均值聚类, 有效检测发生聚类概念漂移和没有发生聚类概念漂移的情况, 从而实现了动态数据的在线聚类, 避免了经典 k-均值方法在动态数据中每次都要对全部数据重新聚类而导致算法速度过慢的问题. 标准数据集和人工社会网络数据集上的实验结果表明, 与经典 k-均值聚类方法相比, 本文提出的动态 k-均值聚类方法能快速高效地处理动态数据聚类问题, 并有效地检测动态数据聚类过程中所产生的概念漂移问题.

关键词: K-均值聚类; 动态 k-均值算法; 动态数据; 概念漂移

Research and Realization of a Web Information Extraction and Knowledge Presentation System

HU Wei

(Experimental Teaching Center, Shanxi University of Finance and Economics, Taiyuan 030006, China)

Abstract: This paper presents an improved dynamical k-means clustering model to solve the dynamical problem, called Dynamical K-means algorithm, in order to solve the problem that only solving the constant clustering problems of classical k-means clustering method. Based on classical k-means method, by analysis and solving the new adding samples of dynamical training data set, local renew or global clustering is performed by the changing range of objective function, and the dynamical data are clustered online. The speed of classical k-means algorithm is slow by the reiterative clustering is needed of every online clustering step, but the speed of Dynamical K-means algorithm is accelerated. Simulation results on standard and artificial social network datasets demonstrate that comparing with classical k-means clustering means, the excellent clustering results can be obtained by this method and the concept drifting phenomenon can be monitored efficiently.

Key words: K-means clustering; dynamical K-means algorithm; dynamical data; concept drifting

随着信息技术尤其是数据库技术的不断发展, 人们能够以更快捷、容易、廉价的方式获取和存储数据, 这使得大量的数据在诸多领域存储下来. 有资料显示, 2011 年全球数据存储量就达到 1.8ZB, 预计 2020 年将增长 50 倍[1]. 然而, 人们的各项活动都是基于知识和智慧的, 如果数据没有经过分析、加工、处理和精炼, 那么数据本身没有多大意义. 为了能够从海量数据中提炼出有用的知识, 人们迫切需要一种能够智能地、自动地把数据转换成为有用信息和知识的技术、方法或工具, 这就导致了数据挖掘技术的产生. 聚类作为一种典型的数据挖掘方法, 一直以来都是人工智能领域的一个研究热点, 被广泛地应用于人脸图像识别、股票分析预测、搜索引擎、生物信息学等领域中.

聚类分析, 就是将物理或抽象对象的集合分组成为由类似对象组成的多个簇的过程. 一般地, 在聚类结果中, 同一类别中的对象有较大的相似性, 不同类别的对象有较大的差异性. 聚类分析的目标就是在相似度量的基础上对数据进行划分. 聚类来源于很多学

① 收稿时间:2012-10-22;收到修改稿时间:2012-12-01

科领域, 包括: 数学、计算机科学、统计学、生物学和经济学等. 在不同领域中, 都有适用于该领域的聚类技术, 并被用来衡量数据之间的相似性[2].

目前, 常用的聚类方法包括: 划分聚类、层次聚类、密度聚类、网格聚类等. 划分聚类就是对给定的包含 n 个对象的数据集, 采用划分方法分为 k 个组 ($k \le n$), 每组代表一个类, 且每个分组至少包含一个数据纪录, 每一个数据纪录属于且仅属于一个分组. 常见的算法如: K-means 算法[3]、K-medois 算法[4]和 CLARANS 算法[5]等. 层次聚类就是对给定的数据集进行层次分解或者合并, 直到所有的记录组成一个分组或者某个条件满足为止, 具体又可分为合并(自底向上)和分解(自顶向下)两种方案, 常见的算法有: Birch 算法、Cure 算法等[6,7]. 基于密度的聚类克服了传统基于距离的聚类算法只能发现超球形类的缺点, 该方法的基本思想是只要一个区域中点的密度大于某个阀值, 就把它加到与之相近的聚类中去, 其代表算法有 DBSCAN 算法[8]、OPTICS 算法[9]等. 基于网格的聚类方法把对象空间量化为有限数目的网格单元, 形成一个网格结构, 所有的聚类操作都在这个量化空间进行, 常见的网格聚类算法有 STING 算法[10]和 CLIQUE 算法[11]. 此外, 基于模型的聚类[12]和基于谱的聚类[13,14]也被众多学者所研究.

在众多的聚类方法中, k-means 方法是一种最经典的应用最为广泛的聚类方法[15,16]. 该方法以各类样本的质心代表该类进行不断迭代, 但该方法只适用于数值型属性数据, 对超球形和凸形数据有较好的聚类效果. 但是, 经典 k-means 算法只能处理静态数据的聚类问题, 对于数据不断发生变化特别是包含概念漂移现象的动态聚类问题, 当数据集更新时, 经典 k-means 方法需要重新对更新后的整个数据集进行聚类, 因此, 对于动态数据的聚类问题, 每当数据更新一次时, 算法就相当于在整个工作集上重新执行一次, 其执行效率是比较低的.

针对传统 k-means 方法不能有效处理动态数据聚类的问题, 本文在经典 k-means 聚类方法的基础上, 通过对新加入的样本与已进行的聚类之间的关系进行分析, 有效地检测发生聚类概念漂移和未发生聚类概念漂移的情况, 对于少数新加入样本后可能产生聚类概念漂移的样本采用经典 k-means 聚类方法重新聚类, 而对于大多数未产生聚类概念漂移的样本所属类别只需要一个简单的更新, 而不需要在整个数据集上重新聚类, 大大增加了新加入样本时聚类的执行效率, 并且能够有效检测动态聚类过程中所发生的概念漂移现象.

1 经典k-means聚类方法

设数据集 $X=\{x_i\}_{i=1}^{n}$ 且 $x_i \in R^d$, k 为指定的聚类个数. K-means 算法的基本工作过程为: 首先, 从 n 个数据对象中随机选择 k 个对象作为初始聚类中心, 其它对象则根据其与已得到的聚类中心的相似度分别分配到最相似的类中. 计算相似度的公式如下, 假设 c_j 为第 j 个类的类中心, 则 x_i 与第 j 类的相似度定义为:

$$s(x_i, Class_j)=\frac{1}{d(x_i,c_j)}=1\Big/\sqrt{\sum_{k=1}^{d}(x_{ik}-c_{jk})^2} \tag{1}$$

然后, 计算每个更新类的新聚类中心, 假设第 j 类中的样本为 $\{x_{j1}, x_{j2}, \cdots, x_{jn_j}\}$, 即包含 n_j 个样本, 该类的聚类中心为 $c_j=\{c_j^p\}_{p=1}^{d}$, 其中, c_j^p 为类中心 c_j 的第 p 个属性, 则聚类中心第 p 个属性分量的求法如下:

$$c_j^p=\sum_{q=1}^{n_j}x_{jq}^p\Big/n_j \tag{2}$$

不断循环迭代, 直到标准测度函数收敛为止(从表现形式上看即更新后的聚类中心与更新前一致), 一般采用均方差作为标准测度函数, 其形式为:

$$J=\sqrt{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-c_i)^2}\Big/(n-1) \tag{3}$$

最终得到的聚类结果中聚类内部应当尽可能紧凑, 不同类类间应当尽可能分离.

但是, 经典 k-means 聚类方法由于需要在每次聚类时衡量所有样本到每个类中心的相似度, 因此算法复杂度较高, 运行速度较慢, 若针对动态数据, 每次都在整个数据集上进行聚类显然是不现实的. 因此, 针对 k-means 方法不能有效处理动态数据聚类的问题, 本文提出一种能够有效处理动态数据聚类问题的改进的动态 k-means 聚类方法, 该方法通过不断地衡量新来样本与已经产生的聚类之间的关系, 仅仅对与新样本有关的类别进行简单更新, 而不需要在整个更新后的数据集上重新进行聚类, 从而大幅度地提高了 k-means 处理动态数据聚类问题的执行效率, 并有效检测动态聚类过程中的概念漂移现象.

2　动态k-means聚类方法

设数据集 $X=\{x_i\}_{i=1}^{n}$ 且 $x_i \in R^d$，k 为指定聚类个数. 由于经典 k-means 方法只能对静态数据集(即数据集本身在聚类过程中不发生变化)的问题进行有效聚类, 而对于动态数据集(如聚类过程中数据逐个增加)的聚类问题, 传统 k-means 方法必须每次都重新对更新后的整个数据集重新聚类. 由第二节知, 经典 k-means 聚类方法处理静态数据的复杂度为 $O(nkl)$, 其中 n 为聚类的样本规模, k 为聚类个数, l 为迭代次数. 采用经典k-means方法处理动态数据, 其复杂度为 $O(nklm)$, 其中 m 为更新的样本数. 若更新较多时, 其复杂度是非常大的. 针对这个问题, 本文提出了一种改进的动态 k-means 聚类方法(Dynamical k-means)用于处理动态数据的聚类问题, 该方法首先对原始的样本集进行初始聚类, 得到一个初始的聚类模式. 当有新样本进入时, 通过衡量该样本与已经得到的不同类之间的相似性关系, 来衡量是否发生聚类概念漂移, 若没有发生概念漂移, 则将其划分到相似度大的类中, 而不需要重新在整个更新的数据集上进行聚类, 否则需要在整个数据集上重新进行聚类, 从而得到更新后的训练结果, 如此反复迭代直到没有新的数据加入或者聚类模式不再发生改变为止.

动态 k-means 聚类算法:

Step1: 选择包含 n 个数据对象的样本集 $X=\{x_i\}_{i=1}^{n}$ (其中 $x_i \in R^d$)进行初始的 k-means 聚类, 得到初始聚类模式 $X \to \{X_j\}_{j=1}^{k}$. 具体方法如下:

Step1.1: 设置初始聚类个数参数 k, 初始化聚类目标函数 $J_1^{(0)}=0$, 初始化聚类中心集 $C=\phi$, 随机选择 k 个样本作为初始聚类中心;

Step1.2: 根据式(1)计算每个样本 x_i ($i=1,\cdots,n$)与所有不同的类 $Class_j (j=1,\cdots,k)$ 之间的相似度, 并将 x_i 归为与其最相似的类中心所属的类;

Step1.3: 根据式(2)计算样本重新分配后的每个类的中心 $C=\{c_j^{(t)}\}_{j=1}^{k}$;

Step1.4: 根据式(3)计算当前的目标函数, 并计算 $r_1^{(t)}=J_1^{(t)}-J_1^{(t-1)}$, 若 $r_1^{(t)}$ 为 0 或者达到迭代步骤的上限值时, 则初始聚类结束; 否则更新 t 值, 并返回 Step1.2 反复迭代执行, 直到 $r_1^{(t)}$ 为 0 时聚类结束, t 就是本次聚类需要迭代的次数.

Step2: 设第 l 次更新数据集时新添加的对象为 x_{n+l}, 对新的数据集进行聚类. 根据式(1)计算 x_{n+l} 与初始得到的 k 个类的相似度 $s(x_{n+l}, Class_j)$, $j=1,2,\cdots,k$.

Step3: 取 $s(x_{n+l}, Class_j)$ 中所对应相似度的最大值, 并根据式(4)将 x_{n+l} 入该最大值所对应的第 j 类当中:

$$X_j = X_j \bigcup \{x_{n+l} \mid s(x_{n+l}, c_j) \geq s(x_{n+l}, c_i), j \neq i\} \quad (4)$$

同时, 根据式(5)计算更新类的新的类心, 其第 p 个属性分量为

$$c_j^p = \left(\sum_{q=1}^{n_j} x_{jq}^p + x_{n+l}^p\right) \Big/ (n_j+1) \quad (5)$$

Step4: 根据式(3)计算当前的目标函数 J_l, 并与上次得到的目标函数 J_{l-1} 进行比较, 设 $\Delta = J_l - J_{l-1}$, 如果 $\Delta > \varepsilon$, 则发生聚类概念漂移, 对新添加后的整个数据集进行经典 k-means 聚类; 反之, 若 $\Delta \leq \varepsilon$, 则没发生聚类概念漂移, 执行 Step5. 在实际问题中, 参数 ε 可根据情况自行选取.

Step5: 观察是否有样本继续更新, 若有, 则返回 Step2 继续执行, 若没有, 则输出最终聚类结果.

算法结束.

3　实验结果及分析

为验证本文提出的改进动态 k-means 聚类方法的有效性, 在四个标准数据集和一个模拟的社会网络数据集上进行了实验, 并将聚类结果与经典 k-means 方法所得到的聚类结果进行了比较. 实验在 1 台 PC 机(2.66Ghz CPU, 1G 内存)上进行测试, 实验平台是 Matlab7.0.

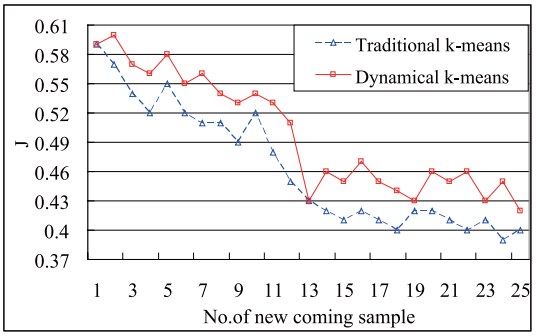
实验采用的标准数据集见表 1, 这些数据集可从网站 http://www.ics.uci.edu/~mlearn/MLRepository.html 上下载. 实验中, 所有数据集进行 10 等分(Spambase 数据集第一份比其它份多 1 个样本, 即第一份包含 174 个样本), 其中 9 份用于初始 k-means 聚类, 剩余 1 份用于数据集的动态变化的新增样本集.

表 1　标准数据集

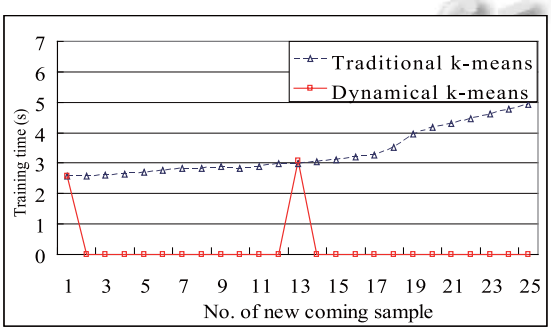
数据集	训练集规模	新增样本规模	数据维数
ASL	250	25	22
Breast_cancer	1900	190	9
Spambase	1731	173	57
Flare_solar	3330	333	9

设置初始聚类个数均为 20, 四个标准数据集分别通过 25、200、173、333 次数据更新, 参数 ε 取每次更新前目标函数的 1/10, 图 1 给出了本文提出的动态

k-means 聚类方法及经典 k-means 聚类方法在 ASL 数据集上每一次新样本加入后所需要的 J 值和训练时间随着新样本加入而得到的变化趋势图.



(a) J 值



(b) 聚类时间

图 1 ASL 数据集

从图 1 可以看出, 在标准数据集 ASL 上, 尽管动态 k-means 聚类方法的 J 值稍许偏高, 但其聚类时间缩短了很多. 此外, 动态 k-means 方法能准确地检测出有 1 个新增样本加入时发生了聚类概念漂移. 由于当新增样本没有被检测为聚类概念漂移点时, 只需要简单地把该样本归于最相似的类别, 而不需要进行全局的 k-means 聚类, 因此, 采用动态 k-means 的方法多数数据加入时聚类时间均近似为 0.

在其他数据集上同样做了相同的实验, 当新增样本较多时, 统计图无法清晰地看出样本变化的情况, 因此采用以下几个指标衡量算法效果. 设新增样本集为 X, 每个数据集中实际发生概念漂移的样本集合为 X_1, 其规模为 n_1, 本文方法检测到的聚类概念漂移样本集合为 X_2, 其规模为 n_2, 概念漂移样本检测率为 p_1、错检率为 p_2、传统 k-means 方法整个过程的训练时间 t_1 与动态 k-means 方法整个过程的训练时间 t_2. 其中, 检测率 p_1、错检率 p_2 定义如下:

$$p_1 = \frac{|X_1 \cap X_2|}{n_1} \tag{6}$$

$$p_2 = \frac{|X_2 - (X_1 \cap X_2)|}{|X|} \tag{7}$$

其他标准数据集上的实验结果见表 2.

表 2 各标准数据集上得到的实验结果

Datasets	n_1	n_2	p_1 (%)	p_2 (%)	t_1 (s)	t_2 (s)
ASL	1	1	100	0	83.4	5.65
Breast_cancer	4	7	100	1.58	573	19.3
Spambase	6	8	83.3	1.73	509	24.6
Flare_solar	11	29	100	5.41	134	48.9

从表 2 中可以看出, 除标准数据集 Spambase 之外, 其他标准数据集中实际发生概念漂移的样本点的检测率都为 100%, 标准数据集 Spambase 的检测率也在 80%以上, 也就是说对于标准数据集 Spambase, 本文提出的动态 k-means 聚类方法只有一个聚类概念漂移的新增样本没有被检测到; 除标准数据集 Flare_solar 之外, 其他标准数据集的概念漂移样本误检率都低于 2%, 尽管标准数据集 Flare_solar 的误检率达到了 5.41%, 但相对于大规模的新增样本集来说, 发生误检的样本依然是较少的; 由于动态 k-means 聚类方法只有少数的被检测为聚类概念漂移的样本进入时, 才对整个数据集进行一次 k-means 聚类, 而其他情况下只根据其与已有类别的相似性来进行划分, 因此训练效率得到了大幅度提高.

此外, 本文还将动态 k-means 方法应用于人工模拟的网络社团结构数据集中, 并与传统 k-means 方法进行了对比. 该人工模拟的网络社团机构数据集包含 100 个顶点, 它们被划分成相同大小的 4 个社团结构, 每个社团结构包含 25 个顶点, 且所有顶点的度均符合如下条件:

$$z_{in} + z_{out} = 15 \tag{8}$$

其中 z_{in} 为每个顶点随机指向落在相同社团中顶点的有向边数目, z_{out} 为每个顶点随机指向落在不同社团中顶点的有向边数目. 图 2 为网络中 100 个顶点在 z_{in}=12, z_{out}=3 时被划分成 4 个社团结构时的示意图. 实验中, 将前 70 个数据作为初始样本集, 后 30 个数据作为动态变化的新增样本集. 实验结果所测得到的 J 值和训练时间如图 3 所示.

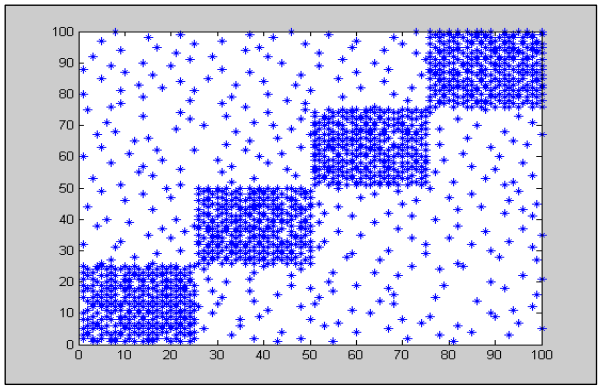
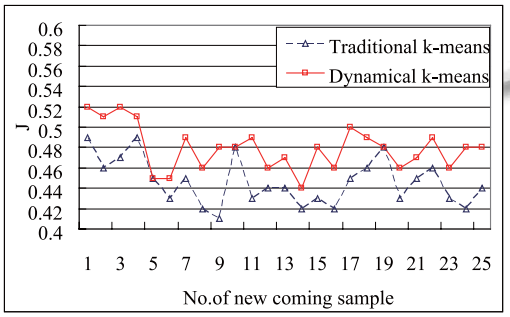
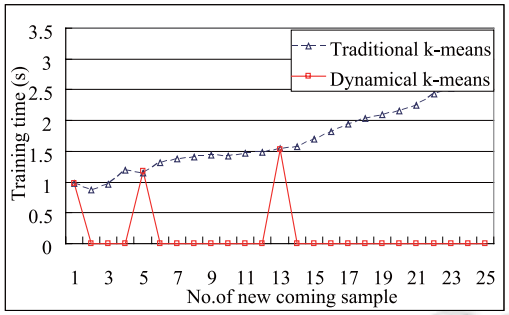


图 2　人工模拟网络社团结构数据集



(a)　J 值



(b)　训练时间

图 3　网络社团结构数据集实验结果

从以上实验结果可以看出，在构造的网络社团结构数据集上，尽管动态 k-means 聚类方法的 J 值稍许偏高，但其聚类时间缩短了很多. 此外，动态 k-means 方法检测了两个新加入的样本发生了聚类概念漂移，其中第 5 个样本点是正常概念漂移点，而第 13 个样本点是错误检测的概念漂移样本点. 由于当新增样本没有被检测为聚类概念漂移点时，只需要简单地把该样本归于最相似的类别，而不需要进行全局的 k-means 聚类，因此，这个过程所耗费的时间忽略不计.

综上可看出，本文提出的动态 k-means 方法能够以较高的训练效率处理在线聚类问题，并有效检测在线学习过程中发生的概念漂移问题，从而及时地更正模型.

4　结语

K-means 方法是目前机器学习领域常用的一种有效聚类方法，本文针对经典 k-means 聚类方法对动态数据聚类的问题需要每次都进行整个数据集的聚类而导致聚类效率低下的问题，提出了一种专门用来处理动态数据聚类的动态 k-means 聚类方法，该方法通过衡量新添加样本与已有聚类模式之间的关系，根据目标函数改变情况决定进行简单地选择最相似的类别进行局部更新还是重新进行全局经典 k-means 聚类，以大幅度提高算法的执行效率，并有效检测发生概念漂移的问题. 在未来的工作中，将进一步探索加速 k-means 聚类方法的动态聚类机制，并与 PCA 等特征选择方法相结合，使其能够被应用到诸如在线图像自动处理、在线网页自动分类等高维的实际动态聚类问题中.

参考文献

1 http://www.zdnet.com.cn/files/mail_con.php?mid=1735,2011,7.

2 Jain AK,Murty MN,Flynn PJ.Data clustering:a review.ACM Computing Surveys,1999,31(3):264–323.

3 MacQueen J.Some methods for classification and analysis of multivariate observations.Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability,Berkeley,1967,1:281–297.

4 Kaufman L,Peter JR. Finding groups in data:an introduction to cluster analysis.Washington:John Wiley & Sons,1990.

5 Ng RT,Han JW.Efficient and effective clustering methods for spatial data mining.Proceedings of the 20th International Conference on Very Large Data Bases (VLDB1994),Santiago, 1994:144–145.

6 Cilibrasi RL,Vitányi PM.A fast quartet tree heuristic for hierarchical clustering.Pattern recognition,2011,44(3):662–677.

7 白旭,靳志军. K-中心点聚类算法优化模型的仿真研究.计算机仿真,2011,28(1):218–221.

8 Ester M,Kriegel HP,Sander J.A density-based algorithm for

discovering clusters in large spatial databases with noise. Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD1996),Portland,Oregon, 1996:125−138.

9 武佳薇,李雄飞,孙涛等.邻域平衡密度聚类算法.计算机研究与发展,2010,47(6):1044−1052.

10 Su MC,Chou CH.A modified version of the k-means algorithm with distance based on cluster symmetry.IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001,23(6):674−680.

11 Agrawal R,Gehrke J,Gunopulos D,et al.Automatic subspace clustering of high dimensional data for data mining application.Proceedings of the ACM SIGMOD Conference on Management of Data (SIGMOD1998),Seattle,1998:94−104.

12 李凯,李昆仑,崔丽娟. 模糊聚类在集成学习中的应用研究.计算机研究与发展,2007,44(z2):203−207.

13 王玲,薄列峰,焦李成.密度敏感的半监督谱聚类.软件学报,2007,18(10):2412−2422.

14 王会青,陈俊杰,郭凯.遗传优化的谱聚类方法研究.计算机工程与应用,2011,47(14):143−145.

15 Kanungo T,Mount DM.A local search approximation algorithm for k-means clustering.Computational Geometry, 2004,28(2/3):89−112.

16 Elkan C.Using the triangle inequality to accelerate k-means. Proceedings of the Twentieth International Conference on Machine Learning(ICML-2003),Menlo Park,AAAI Press, 2003:147−153.

(上接第 106 页)

统定位性能相较于原 LANDMARC 系统也有很好地改善, 系统算法的复杂度得到了明显的减小. 从而, 验证了本文提出的改进方案以及算法的改进思路是可行的, 并且改进优化后的系统对于定位精度和定位性能有较高需求的应用场景具有现实的应用价值及意义.

参考文献

1 J.Hightower and G.Borriello,A survey and taxonomy of location sensing systems for ubiquitous computing,CSE 01-08-03, University of Washington,Department of Computer Science and Engineering,Seattle,WA,Aug.2001.

2 Ni L M,Liu Yunhao,Lau Y C,et al.LANDMARC:Indoor Loca- tion sensing Using Active RFID.Wireless Networks,2004,10 (6):701−710.

3 孙瑜,范平志.射频识别技术及其在定位中的应用.计算机应用,2005,6(5):1205−1208.

4 王远哲,毛陆虹,刘辉,肖基诰.基于参考标签的射频识别定位算法研究与应用.通信学报,2010,31(2):86−92.

5 邓辉舫,马启平,周尚伟.使用无线射频识别(RFID)技术进行室内定位.计算机应用,2008,28(7):1858−1865.

6 韩下林,赵卫东,季军,卫岗,柳先辉.基于 RFID 的室内定位算法及其改进[A].计算机工程与应用,2008,45(11):47−69.

7 周艳,李海成.基于 RSSI 无线传感器网络空间定位算法.通信学报,2009,30(6):75−79.

8 中国射频识别(RFID)技术政策白皮书[P].北京:中华人民共和国科学技术部等十五部委,2006.