

基于数据场的无线传感网数据双重聚类算法

石美红, 赵 辉, 贾郑磊, 雷 燕, 张祥俊
(西安工程大学 计算机科学学院, 陕西 西安 710048)

摘 要: 为了解决大规模高维 WSN 数据的双重聚类问题, 提出一种基于数据场的 WSN 数据聚类方法. 借用数据场的思想, 将 WSN 数据空间非线性映射到势能空间; 结合 WSN 数据的分布特点, 采用概率熵度量数据的质量; 根据数据场形成的等势线分布特征, 采用极小判定法得到数据聚类结果. 通过人工合成数据集实验测试, 结果表明, 对于随机分布的大规模数据集的聚类效果和聚类精确度, 此算法优于 ICC 和 DFCM 的双重聚类算法, 且具有较低的时间复杂度和良好的可扩展性.

关键词: 无线传感器网; 大规模高维数据; 双重聚类; 数据场; 势能

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2016)02-0049-05

An Dual Clustering Algorithm for Data in Wireless Sensor Networks Based on Data-field Model

SHI Mei-hong, ZHAO Hui, JIA Zheng-lei, LEI Yan, ZHANG Xiang-jun

(School of Computer Science, Xi'an Polytechnic University, Xi'an 710048, China)

Abstract: The paper has proposed a novel self-organizing-mapping algorithm based on data-field model for dual clustering of large-scale and high dimensional data in Wireless Sensor Networks (WSNs) in order to has good performance in cluster problems. The method maps the WSNs data from data space to the appropriate potential space in data field, which measures the interactions of the elements in large-scale and high dimensional data by taking probabilistic entropy of data distribution in the WSNs as the mass of data field, thus generating a two-dimension data field. Then, by employing distribution features of the potential center and the equipotential lines, without significantly increasing the time complexity, the good clustering result is obtained by minimum potential difference determination method. The comparing experiments on the synthetic datasets demonstrate the effectiveness of the algorithm. Experimental results show that the proposed method improves the clustering effect and has exact clustering result compared with other dual clustering algorithm, i. e. ICC and DFCM, and it has good scalability.

Key words: wireless sensor networks; large-scale and high dimensional data; dual clustering; data-field; potential

1 引言

作为物联网的核心—无线传感器网络 (Wireless Sensor Networks, WSN) 因其硬件和应用的灵活性和多样性, 已被广泛地运用于各种领域, 如何减少节点能量消耗、延长网络寿命是 WSN 应用面临的一个实

用性问题^[1].

人们采用数据聚类法消除冗余数据, 是减轻网络负担, 节约节点能耗的重要方法. 文献[2]首次研究了双重聚类思想, 提出了一种基于聚类和分类交错的 ICC 双重聚类算法, 取得了较好的聚类效果, 但计算复杂度较高, 且用 SVM 解决大规模高维节点数

收稿日期: 2015-04-17; 修回日期: 2015-06-15

基金项目: 国家科技支撑计划项目 (2014BAF07B01)

据的多分类问题存在一定的局限性. 随后, 人们又提出了其他改进的双重聚类算法, 比如 Huang 等人^[3]提出 DFCM 双重聚类算法. 目前, 从场论的角度, 提出了一种基于数据场的数据处理方法, 而被广泛应用^[4-5]. 从理论和实验分析验证了该方法的可行性. 鉴于此, 本文提出了一种基于数据场的 WSN 数据双重聚类算法 (Dual Clustering Algorithm Based on Data-Field, DCDF).

2 数据场聚类模型分析

数据场思想认为, 数据空间内每个对象都是具有一定质量的质点, 它们不是孤立的, 位于场内的任何对象都会受到其他对象的联合影响或作用而产生势能 (或势值), 具有相同势能的对象组成一个等势线 (或面), 由等势线围绕形成的中心称为势心, 其势能为局部极大值; 环绕势心的邻域等势线分布越密集, 其势能就越大; 以势心为中心的等势线分布构成了势能的特征空间.

为了直观地表达数据场聚类的特性, 图 1 给出了可视化的数据场聚类模型. 图 1(a) 是具有 3 类的人工合成数据的空间分布, 将其分别映射到二维和三维的势能空间后的分布图如图 1(b)、(c) 所示.

由图 1(b)、(c) 不难看出, 势心所在位置是当前局部数据簇的中心, 其势能为局部数据簇的极大值; 距势心近且分布密集的等势线 (或面) 势能比距势心远且分布稀疏的等势线 (或面) 势能要高. 在数据场中, 各数据簇是以势心为中心, 以环绕周围邻域的等势线 (或面) 形成自然抱团特性. 从聚类的角度, 势心反映了数据簇的聚类重心; 环绕势心的邻域等势线分布反映了各数据簇的自然聚类. 因此, 通过对数据场形成的环绕势心向外扩散的等势线的合理划分, 实现数据的自组织聚类.

3 WSN 数据场的表示与特性

WSN 数据空间是由空间位置随机分布的节点, 通过某种途径或手段, 对所在位置的物体 (如温度、湿度、气体等) 进行观测所获得的动态分布数据集.

定义 1 WSN 数据空间: 假设: $G = \{(x, y) \mid x \in [0, W-1] \wedge y \in [0, H-1] \wedge x, y \in S\}$ 是二维节点构成的有限空间, $D = \{d_i^G \mid d_i^G \in R \wedge i = 1, 2, \dots, N\}$ 是由在 G 空间上节点所感知的数据集, 其中, W 和 H 分别表示 WSN 在 G 空间上的宽、高; $d_i^{(x,y)} = (d_{i1}^{(x,y)}, d_{i2}^{(x,y)}, \dots, d_{ip}^{(x,y)})$ 是在 G 空间上位于 (x, y) 的第 i 节点感知的 p 维数据; N 是 WSN 中节点数目; R

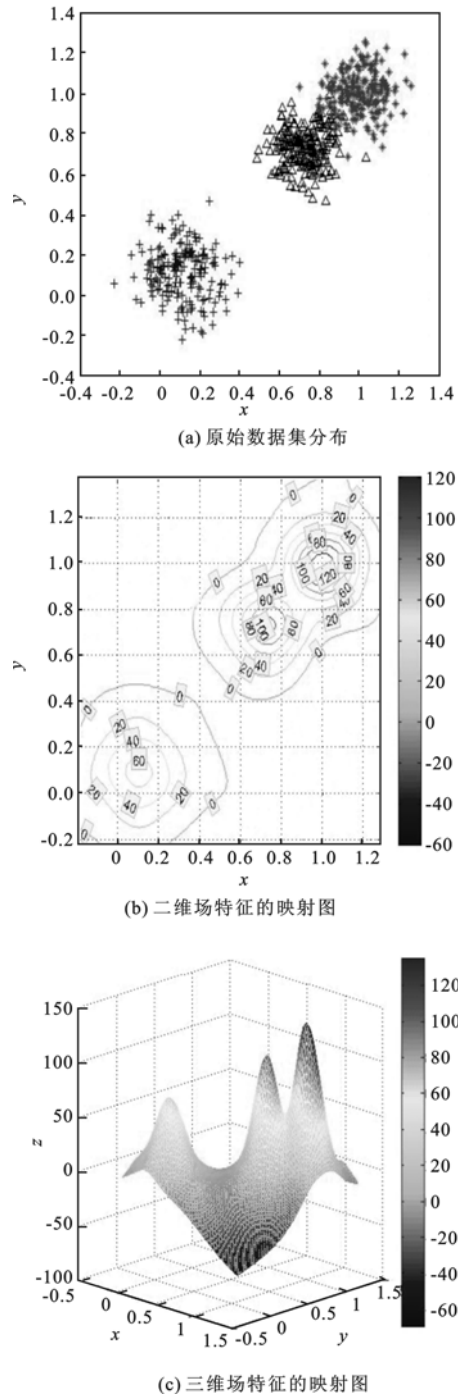


图 1 数据场聚类模型

为实数, 则 WSN 数据空间由节点的二元组 $U = (G, D)$ 确定.

定义 2 WSN 数据场: 给定 WSN 数据空间 $U = (G, D)$, 对于 G 在空间内位于 (x, y) 任意上的第 i 节点, 视为带有质量为 $m_i^{(x,y)}$ 的质点, 其周围存在一个来自位于任意 (k, l) 上其他 j 节点的联合作用场, 其势能大小为

$$\Phi(x, y) = \sum_{\substack{k \in [0, W-1] \\ l \in [0, H-1] \\ j \in [1, N]}} \varphi_j^{(k, l)}((x, y), d_i^{(x, y)}) \quad (1)$$

式中, $\varphi_j^{(k, l)}((x, y), d_i^{(x, y)})$ 表示位于 (k, l) 的第 j 节点对位于 (x, y) 的 i 节点所产生的势能贡献。由此形成的场称为 WSN 数据场。

数据场的核心在于势函数的定义,其选取并不唯一。经分析,文献[6]认为拟核力场的势函数形态数学性质更好、物理意义更合理。因此,结合聚类目标,本文采用拟核力场的高斯势函数,其定义式为 $\varphi_j^{(k, l)}((x, y), d_i^{(x, y)}) = m_j^{(k, l)} \cdot$

$$\exp\left(-\frac{\|(x, y), (k, l)\|^a}{2\sigma^a}\right) \quad (2)$$

式中, $m_j^{(k, l)}$ 表示位于 (k, l) 的第 j 节点的质量; $\|(x, y), (k, l)\|$ 为两节点间距离; $\sigma \in (0, +\infty)$ 为控制质点间相互作用力程的影响因子; a 为距离指数,一般取 2。为了提高算法执行效率,根据高斯函数的 3 σ 性质,仅考虑距离不超过 $\xi = [3\sigma/\sqrt{2}]$ 的影响范围。

由上述分析不难看出,质量的选取是影响 WSN 数据聚类性能的关键因素之一,不同的质量表示突显出不同的质点特征。概率分布是度量分布状态随机性测度,熵是度量信息状态不确定性的测度。针对 WSN 数据空间分布的特点,选用概率熵度量 WSN 数据场的质点质量。首先将 WSN 节点的各属性数据进行无量纲化处理至区间 $[0, 1]$, 以消除各维属性数据域分布的不一致性,再利用式(3)计算场内位于 (x, y) 的第 i 节点质量:

$$m_i^{(x, y)} = \sum_{j=1}^p \omega_j \cdot d_{ij}^{(x, y)} \quad (3)$$

式中, $d_{ij}^{(x, y)}$ 表示位于 (x, y) 的第 i 节点的第 j 维属性数据; $\omega_j = \frac{1 - H_j}{\sum_{t=1}^p (1 - H_t)}$ 是第 j 维属性数据在 WSN

数据场中的分布权重, $0 \leq \omega_j \leq 1$, $\sum_{j=1}^p \omega_j = 1$; $H_j = -$

$\frac{1}{\ln N} \sum_{i=1}^N P_{ij} \ln P_{ij}$ 为第 j 维属性数据在 WSN 数据空间内的平均熵; $P_{ij} = d_{ij} / \sum_{h=1}^N d_{ih}$ 表示第 i 节点的第 j 维属性数据在 WSN 数据空间中的分布概率。

通过 WSN 数据空间中节点数据的全局分布信息特征和邻域的相互关系,建立 WSN 数据场内质点间产生的势能差异度量;通过质点间相互作用形成的等势线分布特征和邻域的相互关系,建立 WSN 数据空间自然聚类的度量。

性质 1 WSN 数据场势能的影响因素:包含质点的质量、质点间的距离和影响因子 σ 等影响因素,其相互关系是:(1)在相同的距离内,质量越大,对应节点数据的概率熵越小,其分布越有序,产生的势能就越大,成为聚类重心的可能性就越大;(2)当质量相同时,质点间的相对距离越小,对应节点数据的分布越密集,彼此间的相互作用力就越大,其势能也就越大,形成自然抱团的特性就越明显;(3)当质量和质点间距离不变时,影响因子 σ 越大,控制质点间的相互作用力就越强,彼此间产生的势能影响就越大,场内所有质点的势能差异也就越小当 $\sigma \rightarrow \infty$ 的极端情况下,各质点势能趋于相同。

性质 2 WSN 数据场的势能特性:(1)数据场势能具有上下限 $[0, \sum_{\substack{k \in [0, W-1] \\ l \in [0, H-1] \\ i \in [1, N]}} m_i^{(k, l)}]$;(2)具有同类属性

的数据,对应于数据场的质点质量相同,若质点间距离相近,表明它们的空间分布密集,彼此间产生的作用力就强,对应的平均势能高,距场势心就越近,抱团特性也越明显,当质点间距离为零的极端情况下,它们将位于同一等势线或势心;若质点间相距较远,表明它们的空间分布稀疏,彼此间产生的作用力弱,对应的平均势能偏低,距场势心就越远,当距离超出一定范围时,因势能分布的模糊性,将可能位于不同簇类间的重叠区;(3)具有不同类属性的数据,对应于数据场的质点质量不同,若质点间距离较远,那么彼此间产生的作用力较弱,当质点间距离为无穷的极端情况下,彼此间无作用力影响,因质量不同致使它们分布在不同的局部邻域数据簇中;若质点间相距较近,那么因大质量质点产生的势能大于小质量质点,导致它们间存在着势能差,在局部邻域等势线表现出阶跃状的不连续性,当相距很近时,因势能分布的不确定性,将可能位于不同簇类间的重叠区。

4 算法描述

基于 WSN 数据场的数据聚类基本思想是,利用等势线环绕形成的势心划分,获得数据簇的分类数;利用环绕势心向外扩散的等势线的分布特征,采用绩效判定法,实现数据聚类。DCDF 聚类算法的具体步骤如下所示。

输入:数据集 U , 辐射因子 σ , 迭代误差 ε , 初始化 $\{\Pi_t | t = 1, 2, 3, \dots, k\}$ 为 0

输出:聚类划分结果 $\{\Pi_1, \Pi_2, \dots, \Pi_k\}$

Step1:初始化, $\{\Pi_t | t = 1, 2, 3, \dots, k\}$ 为 0, 计算 $\xi = 3\sigma/\sqrt{2}$;

Step2: 据式(3)计算场内各质点的质量 $m_i^{(x,y)}$, $x \in [0, W-1], y \in [0, H-1], i = 1, 2, \dots, N$;

Step3: 根据式(1)、(2)计算场内各质点的势能 $\Phi(x, y)$;

Step4: 将势能相等的质点连接形成等势线;

Step5: 势心查找与簇聚类粗划分.

① $t = 1$;

② 计算 $c_t = \text{Max}(\Phi(U - \sum_{k=1}^{t-1} \prod_k))$ 为簇聚类势心;

③ 由 c_t 势心局部向外扩散, 找到局部邻域极小势能的等势线, 并将势心到极小势能等势线内的数据进行类别标记形成簇 $\prod_t$;

④ 当等势线的势能极小或逐渐变大时, 将邻近极小势能等势线的邻域归为类边界, 并开始下一个簇聚类, 处于类边界未标记的数据记为 d_{Spet}^c ;

⑤ 若数据未全部标记完, 则 $t = t + 1$, 返回至②执行, 否则, 找到 t 个簇聚类的势心 $\{c_1, c_2, \dots, c_t\}$, 作为簇聚类的粗划分;

Step6: 计算各簇聚类平均势能 $\overline{\Phi(\prod_k)}$:

$$\overline{\Phi(\prod_k)} = \frac{1}{\text{Num}_k} \cdot \sum_{d_i^{(x,y)} \in \prod_k} \Phi(x, y) \quad (4)$$

式中, $k = 1, 2, \dots, t$, Num_k 表示第 k 簇聚类包含的质点数目;

Step7: 计算标记为 d_{Spet}^c 的类边界数据势能 $\Phi(x, y)$ 与其他各簇聚平均势能 $\overline{\Phi(\prod_k)}$ ($k = 1, 2, \dots, t$) 的距离, 将位于 $G = (x, y)$ 的 $d_{\text{Spet}}^{(x,y)}$ 数据归类为 $\min\{\text{dist}(\Phi(x, y), \overline{\Phi(\prod_k)}) | x \in [0, W-1] \wedge y \in [0, H-1] \wedge k \in [1, t]\}$ (5)

Step8: 按式(4)计算经类边界划分后的每个簇聚类的平均势能 $\overline{\Phi(\prod_k)}$, 判别它与 $\overline{\Phi(\prod_k)}$ 间误差是否小于等于 ε , 若不是, $\overline{\Phi(\prod_k)} = \overline{\Phi(\prod_k)}$, 转到 Step7, 重新计算类边界的划分, 否则, 聚类结束;

Step9: 输出划分聚类结果.

5 实验与结果分析

针对无线传感器网络数据不易获取的特点, 采用人工合成数据集, 在 Intel Core-i3 处理器、4 GB 内存、Win7 64 位操作系统和 Matlab7.6(R2008a) 的实验平台下进行了大量的数据聚类实验, 并与 ICC 算

法和 DFCM 算法进行了对比实验. 为了验证算法的有效性, 分别采用文献[7-8]给出的 DB 和 CA 的聚类质量评价指标. DB 越小, 聚类效果越好; CA 越大, 聚类的精确度就越高.

在合成数据集下, 用可视化图和 DB、CA 聚类性能指标, 对本文算法的聚类结果进行评价. 合成数据集分为均匀分布和非均匀分布两种, 每个数据集含 1 000 个节点, 每个节点具有 2 维位置和 2 维数据的 4 维属性, 其原始合成数据集分布如表 1 第 2 列所示. 分别采用 DCDF 算法、ICC 算法及 DFCM 算法对不同类型数据集进行了聚类, 其结果和聚类性能指标如表 1 所示. 从聚类结果看, 在不同数据分布下, 三种算法都能正确地聚类数据. 但从聚类性能上看, 对于均匀分布的数据集, 本文算法的聚类效果与 DFCM 聚类效果相差不大, 其 DB 指标较为接近, 但 CA 指标略高; ICC 算法因对类边界重叠区的部分数据误划分, 致使 DB 和 CA 的性能不如其他算法. 对于非均匀分布的数据集, 本文算法的聚类性能仍优于 ICC 和 DFCM 的算法.

表 1 不同算法的聚类性能指标对比

聚类方法	指标($K=4$)		指标($K=3$)		指标($K=5$)		
	DB	精确度/ %	DB	精确度/ %	DB	精确度/ %	
本文算法	DCDF	0.160 3	98.4	0.144 8	98.8	0.159 2	97.9
	ICC	0.172 5	96.7	0.159 3	96.7	0.164 3	96.1
	DFCM	0.169 8	97.9	0.175 1	94.5	0.180 5	95.4

为了进一步衡量各算法的扩展性能, 用不同大小的数据集进行了测试, 其结果如图 2 所示, 不难看出, 随着数据集规模的增长, 本文算法的执行时间呈线性变化, 而 ICC 时算法的执行时间迅速递增, 呈非线性变化, DFCM 算法的执行时间介于两者之间.

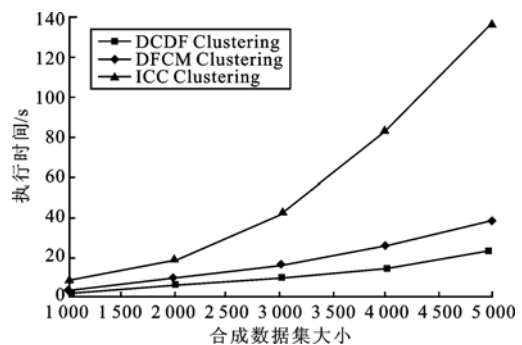


图 2 不同节点数目下算法执行时间比较

6 结束语

为解决随机分布且类边界存在重叠交错的大规模高维 WSN 数据的双重聚类问题,提出了一种基于数据场的 WSN 数据聚类方法. 主要工作如下:(1)借用数据场的思想,构建的聚类模型充分考虑了 WSN 数据的分布特征和双重属性,使之聚类效果和精确度较好,同时,时间复杂度较低;(2)聚类模型对大规模高维数据集的处理具有良好的扩展性. 由于 WSN 数据具有时空的相关性,如何扩展本文算法使之能够处理时变动态数据集有待进一步研究.

参考文献:

- [1] Rault T, Bouabdallah A, Challal Y. Energy efficiency in wireless sensor networks: A top-down survey[J]. Computer Networks, 2014, 14(12): 22737-22753.
- [2] Tai Chihhua, Dai Biru, Chen Mingsyan. Incremental clustering in geography and optimization spaces[J]. Advances in Knowledge Discovery and Data Mining, 2011, 5(4): 707-724.
- [3] Huang Jianghua, Zhang Junying. Distributed dual clustering algorithm based on FCM for sensor streams[J]. Advances in Information Sciences and Service Sciences.

2011, 3(5): 201-209.

- [4] 李文平,杨静,印桂生,等. 数据场典型相关分析及其在图像分割中的应用[J]. 自动化学报, 2015, 41(4): 772-784.
- [5] 邓敏,彭东亮,刘启亮,等. 一种基于场论的层次空间聚类算法[J]. 武汉大学学报:信息科学版, 2011, 36(7): 847-852.
- [6] 李德毅,杜鹂. 不确定性人工智能[M]. 北京:国防工业出版社, 2005.
- [7] Davies D L, Bouldin D W. A cluster separation measure. Pattern analysis and machine intelligence [J]. IEEE Transactions on, 1979, 3(2): 224-227.
- [8] Ahmad A, Dey L. A k-mean clustering algorithm for mixed numeric and categorical data[J]. Data and Knowledge Engineering, 2007, 63(2): 503-527.

作者简介:

石美红 女,(1956-),教授. 研究方向为智能信息处理、图形图像处理与模式识别等. E-mail: meihong_shi@163.com.

赵辉 男,(1990-),硕士研究生. 研究方向为智能信息处理.

贾郑磊 男,(1993-). 研究方向为智能信息处理.

雷燕 女,(1994-). 研究方向为智能信息处理.

张祥俊 男,(1993-). 研究方向为智能信息处理.

(上接第48页)

参考文献:

- [1] 张剑波,卢兰光,李哲. 车用动力电池系统的关键技术与学科前沿[J]. 汽车安全与节能学报, 2012, 3(2): 87-104.
- [2] Lu L, Han X, Li J, et al. A review on the key issues for lithium-ion battery management in electric vehicles [J]. Journal of power sources, 2013, 226(6): 272-288.
- [3] Cheng K W E, Divakar B P, Wu H, et al. Battery-management system (BMS) and SOC development for electrical vehicles [J]. Vehicular Technology, IEEE Transactions on, 2011, 60(1): 76-88.
- [4] 杜祺漳,梁柱扬. 基于"飞电容"技术的动力锂离子电池组保护系统的设计[J]. 电子工程师, 2007, 33(8): 54-56.
- [5] Malcovati P, Brigati S, Francesconi F, et al. Behavioral modeling of switched-capacitor sigma-delta modulators

[J]. Circuits and Systems I: Fundamental Theory and Applications, IEEE Transactions on, 2003, 50(3): 352-364.

- [6] 何乐年,王忆. 模拟集成电路设计与仿真[M]. 北京:科学出版社, 2008: 172-177.

作者简介:

岳泰 男,(1990-),硕士研究生. 研究方向为新能源汽车电池管理系统核心芯片.

乌力吉(通讯作者) 男,(1965-),博士,博士生导师. 研究方向为汽车电子与信息安全.

E-mail: ljiwu@mail.tsinghua.edu.cn

张向民 男,(1966-),硕士,助理研究员. 研究方向为汽车电子与信息安全.

田光宇 男,(1966-),博士,博士生导师. 研究方向为汽车工程.