

传感器网络分布式免疫遗传聚类算法研究

洪月华

(广西经济管理干部学院 计算机系, 广西 南宁 530007)

摘 要: 本文研究无线传感器网络数据的聚类分析问题. 针对传统 k-means 对初始聚类中心敏感和易于陷入局部次优解的缺点, 提出一种基于传感器网络的分布式免疫遗传 k-means 聚类算法. 该算法将聚类中心作为染色体, 通过遗传算法来优化传统 k-means 聚类算法的初始聚类中心, 将免疫算法的选择操作引入染色体的遗传进化中, 使染色体的浓度和适应度共同对其在进化中被选择产生影响, 实现了染色体种群的多样性保持机制和自我调节功能, 将搜索工作引向全局最优, 较好地解决了 k-means 算法的早熟现象问题. 实验结果证明, 本文算法改进了数据的聚类划分效果, 能够把聚类结果快速收敛至全局最优, 聚类准确率高.

关键词: 聚类分析; 免疫算法; 遗传算法; 无线传感器网络

中图分类号: TP393

文献标识码: A

文章编号: 1000-7180(2013)03-0156-04

Research on Distributed Immune Genetic Clustering Algorithm Based on Sensor Network

HONG Yue-hua

(School of Computer, Guangxi Economic Management Cadre College, Nanning 530007, China)

Abstract: The problem of wireless sensor network data clustering analysis is researched. In order to solve the shortcomings that traditional k-means is sensitive for the initial clustering center and easily fall into local optimal solution, a distributed immune genetic algorithm k-means based on sensor network is proposed. In the algorithm, cluster center has been as chromosomes, and through the genetic algorithm to optimize the initial cluster centers of traditional k-means clustering algorithm, meanwhile the selecting operation of immune algorithm is introduced into the genetic evolution of chromosome, causing the concentration and fitness of chromosome will impact whether the chromosome is selected in evolutionary, which can realize chromosome population diversity maintaining mechanism and self regulating function, and the searching achieves the global optimum, so the algorithm can solve the prematurity problem of k-means algorithm better. The experimental results show that, the algorithm has been improved data clustering effect, and made the clustering can converge to the global optimal clustering rapidly, and the clustering has higher accuracy.

Key words: clustering analysis; immune algorithm; genetic algorithm; wireless sensor network

1 引言

聚类是数据挖掘的一个重要研究领域, 是一种重要的无监督学习方法, 其目的是将对象集合分成若干个特征相似的子集合(又叫类或簇), 簇内对象相似性尽可能大, 簇间对象差异性也尽可能大.

无线传感器网络是分布式网络, 而现有的传统聚类算法的工作方式都是集中式聚类, 根本不合适在无线传感器网络中使用. 在 WSN 上进行数据聚类采用传统集中式聚类算法会产生大量的数据通信, 进而消耗节点能量, 最终导致网络寿命降低. 因此鉴于传感器节点的计算能力、存储容量、电源和网

收稿日期: 2012-06-12; **修回日期:** 2012-07-24

基金项目: 广西自然科学基金青年项目(2012jjBAG0074); 广西教育厅科研项目(201010LX428, 201106LX504, 201106LX510, 200103YB195)

络通信能力高度受限的情况,应用在 WSN 上的聚类算法必须是分布式的算法,只有这样才能高效使用能源,使得网络的寿命得到延长。

本文根据经典 k-means 聚类算法的特性,提出一种适用于传感器网络的分布式免疫遗传 k-means 聚类算法。

2 传感器网络分布式 k-means 聚类描述

设网络中用 n 个传感器节点 $J_1, J_2, \dots, J_n$, 来监测数据, 相应用 $X_1, X_2, \dots, X_n$ 表示节点采集的数据, 则整个网络的数据集合 $X = \bigcup_{i=1}^n X_i$ 其中数据 X_i 有 q 个属性即 $X_i = (x_{i1}, x_{i2}, \dots, x_{iq}), (i = 1, 2, \dots, n)$ 。

基于传感器网络的分布式 k-means 聚类算法的核心思想是: sink 节点随机地从传感器网络中抽取 k 个数据作为簇的初始聚类中心 $Z_j (j = 1, 2, \dots, k)$, sink 广播 k 个簇心发送到各个传感器节点。每个节点计算本地的高维冗余数据集合 X_i 中每个数据到这 k 个簇心的最短距离, 把 X_i 划分成 k 个簇, 同时将节点 id, 每个簇中数据个数等簇信息传输到该节点的父节点。而当父节点收到其全部孩子节点的簇信息后, 就将收到的和本地的簇信息统一起来再往上层节点发送, 直至 sink 收到全部孩子节点上送的簇信息, 然后 sink 合并所有簇信息, 接着判断如果聚类目标函数达到最小, 说明聚类结果稳定, sink 就发送停止迭代信息给各个传感器, 否则在 sink 里对每个簇的平均值重新计算得到新簇心, 将其又往下发送到各个传感器节点, 继续做下一次迭代更新^[1]。每次迭代操作都是向着降低簇间相似性, 提高簇内紧凑性的方向发展。

本文的分布式 k-means 聚类算法就是要把传感器网络采集得到的 n 个 q 维数据划分成 k 个簇 $C_j (j = 1, 2, \dots, k)$, 对应 Z_j 是第 j 个簇心, 使得簇内的数据相似度尽可能高, 簇间的数据相似度尽可能底, 同时还满足: $X = \bigcup_{i=1}^k C_i, C_i \neq \emptyset, C_i \cap C_j = \emptyset (i, j = 1, 2, \dots, k; i \neq j)$ 聚类算法的目标函数定义为传感器网络的所有数据到对应簇心距离的总和达到最小(即簇内平方误差和最小):

$$E = \min \left(\sum_{j=1}^k \left(\sum_{i=1}^n \text{flag} \cdot d^2(X_i, Z_j) \right) \right) \quad (1)$$

Flag 是一个变量, 当 X_i 在簇 C_j 里取值 1, 否则取值 0; $d(X_i, Z_j) = \|X_i - Z_j\|$ 是数据对象 X_i 到对应簇 C_j 所在中心 Z_j 的欧式距离。

网络数据进行聚类后鉴于簇内节点数据的相似性很高, 因此仅需在簇内选取一个节点作为代表, 把它的数据输送到基站即可。这样在传感器网络里花销在采集和传送数据的能量就得到了降低, 从而使得网路的寿命得到了延长。

3 免疫遗传 k-means 聚类算法

3.1 算法的基本思想

k-means 算法是一种经典的聚类分析方法, 因其扩展性好、算法简单、能有效处理高维大数据集和收敛速度快而被广泛应用。然而 k-means 算法对初始聚类中心敏感, 即当初始值不同时, 极有可能会得到不同的聚类结果。同时该算法是一种局部搜索技术, 用目标函数进行聚类, 寻找最优解通常都是使用爬山式的梯度下降法, 过早收敛于局部次优解却得不到全局最优解^[2]。

针对上述 k-means 算法的不足, 本文提出用免疫、遗传算法来改进 k-means 聚类算法。用遗传算法进行聚类能较好地优化初始聚类中心, 防止了选择初始聚类中心的盲目性, 由此获得全局最优解, 而将免疫算法所具有的抗体多样性保持机制、自我调节和免疫记忆特性引入遗传算法中, 用免疫系统中基于抗体浓度的选择更新, 替换遗传算法的基于个体适应度的选择复制, 能够有效将搜索工作引向全局最优, 进而避免了 k-means 算法中的早熟现象。

3.2 算法的实现

免疫遗传优化 k-means 算法是参考标准遗传算法的框架结构结合生物机体免疫系统的进化更新机制, 是本文提出的一种新的随机搜索方法。免疫遗传优化 k-means 聚类算法的几个重要步骤是染色体编码、种群初始化、选择、交叉、变异操作。

3.2.1 染色体编码方案

鉴于传感器网络数据集合 X 中的数据个数 n 总是远远比其聚类数目 k 大, 故本文算法使用基于簇中心的浮点数编码方式会更有效。 k 个簇心就构成一条染色体, 即 $Z_1 Z_2 \dots Z_k$ 。由前面传感器数据的组成知 Z_i 也有 q 个属性即 $Z_i = (z_{i1}, z_{i2}, \dots, z_{iq}), (i = 1, 2, \dots, k)$, 则每条染色体的浮点码串的长度是 $k * q$ 。

3.2.2 种群初始化

随机地从传感器网络数据集合 X 中选 k 个数据作为一条初始染色体进行浮点数编码, 重复做 p 次就产生 p 个初始染色体。初始种群就是由这 p 个染色体组成, p 的值不能太大也不能太小, 太大易引起

算法消耗很多时间,太小则种群易失去多样性。

3.2.3 染色体适应度函数

适应度函数作为染色体遗传进化过程中优胜劣汰的主要根据,是评价染色体性能好坏的指标。

最好的聚类划分效果就是同簇内的数据距离越小越好,而簇心间距应尽可能大。综合公式(1),本文选取的适应度函数为:

$$f_t = \frac{\max_{i,j=1}^k \|Z_i - Z_j\|}{E} \quad (2)$$

t 是染色体编号, $t \in [1, p]$ 。当 f_t 越大,目标函数 E 越小,簇心间距越大,聚类效果就越好,说明产生的 k 个簇越紧凑和独立。反之,当 f_t 越小,目标函数 E 越大,簇心间距越小,聚类效果就越差。同时算法在两处用到 E ,节省了在空间和时间上的消耗。

3.2.4 基于免疫算法的选择操作

选择是按照遗传进化中适者生存的原则来开展的操作,适应度越高的染色体得到参加后代繁殖的机会也就越高,使用轮盘赌方法,选择染色体 t 的概率是:

$$p_t = f_t / \sum_{i=1}^p f_i \quad (3)$$

式中, p 种群大小, $t = 1, 2, \dots, p$ 。针对基于标准遗传的聚类算法在前期收敛速度快,而后期当染色体之间的个体差异变小时则收敛速度变慢^[3],本文研究将免疫系统^[4-5]的浓度概念应用于遗传算法的选择操作,以此来调整染色体个体被选中的概率,从而避免上述不足和缺点。

染色体浓度定义如下:

$$den = \frac{\text{种群中染色体相同的条数 } d}{\text{种群大小 } p} \quad (4)$$

则得到这 d 条染色体的个体浓度概率是 $p_{den} = (1 - den)/p$,其它的染色体浓度概率是 $p_{den} = (1 - den + den^2)/p$,全部染色体的浓度概率总和是 1。

综合式(3)和(4),得到选择染色体 t 的最终概率是:

$$p_t = f_t / (p_{den} \times \sum_{i=1}^p f_i) \quad (5)$$

式中, p 和 t 的值同式(3)。

由式(5)知,染色体被选择的概率由适应度和浓度一起决定。因此不仅维持适应度促进选择,算法的收敛速度得到加快;而且经过浓度的调节抑制选择,确保群体中染色体的多样性,防止过早收敛。

3.2.5 自识别交叉

交叉操作是指两个父代染色体交换部分基因而

得到新的子代染色体。交叉是产生新染色体的重要方法,同时对算法的全局搜索能力起到直接的影响作用^[6]。为确保父代染色体的优良性能保留到子代,本文的交叉策略使用自识别交叉,即两个父代染色体是否进行交叉操作由父代间的相似程度来决定。只有当两个父代染色体的相似度 xs 小于阈值 δ 时才能进行交叉。

假设有两个父代染色体 t_1 和 t_2 ,则这两个染色体的相似度定义如下:

$$XS(i, j) = len / (k * q) \quad (6)$$

其中 len 是 t_1 和 t_2 的最长相同子串的长度, $k * q$ 是染色体的浮点码串的编码长度, i 和 j 是染色体中的基因位置序号, $0 < i, j < k * q$, 同时 $i \neq j$ 。

阈值 δ 定义如下:

$$\delta = (1 + \sqrt[3]{g/g_{\max}}) / 3 \quad (7)$$

式中 $\delta \in (1/3, 2/3]$, g 是目前进化代数,而 $g_{\max}$ 是最大进化迭代代数。

3.2.6 自适应变异

变异是一种局部搜索技术,对种群里染色体的多样性有贡献,能避免早熟现象发生。变异操作就是染色体编码串里的基因位上的基因值按变异概率发生随机改变后形成一个新的染色体。本文对基因的变异采用自适应方法^[7]。由编码方案知染色体里的基因 z_{ij} 表示第 i 个簇心的第 j 维值,则变异后该染色体的基因 z'_{ij} 为:

$$z'_{ij} = \begin{cases} \lambda(z_{ij}^{\max} - z_{ij}) + z_{ij}, & \lambda \geq 0 \\ \lambda(z_{ij} - z_{ij}^{\min}) + z_{ij}, & \lambda < 0 \end{cases} \quad (8)$$

式中, $i \in [1, k]$, $j \in [1, q]$, 分别用 $z_{ij}^{\min}$ 和 $z_{ij}^{\max}$ 表示第 i 个簇心的第 j 维的最小和最大值;根据均匀分布原则, λ 应在 $[-(f - f^{\min}) / (f^{\max} - f^{\min}), (f - f^{\min}) / (f^{\max} - f^{\min})]$ 内产生, f 为变异染色体的适应度值, $f^{\min}$ 和 $f^{\max}$ 是当代种群里染色体性能最差和最优的染色体个体适应度值。式(8)使得性能最好的染色体的基因就不会参与变异,反之性能最坏的染色体的基因则变异值较大,这样不但确保了染色体的多样性,而且极大地提高算法的搜索速度和效率。

4 仿真实验和性能比较

仿真实验的配置是 Intel Pentium IV 3.0GHz 的 CPU, 2GB 的 DDR2 内存, Windows XP 操作系统, 使用 Matlab 和 VC++ 编程实现算法。为了检验本文算法的优越性能,将它与传统 k-means 和基于遗传的 k-means 聚类算法对 Iris 数据集进行了聚

类准确率、收敛速度与精度的比较. 实验数据集 Iris 来自文献[8]里三种植物的 150 个样本数据(每种植物各 50 个), 每个样本数据有 4 个属性, 代表植物的四种特征. 相关参数设置: 种群大小 $p=36$, 交叉概率 p_c 和变异概率 p_m 分别取 0.7、0.006, 最大进化迭代代数 $g_{\max}=200$. 分别运行 20 次三种算法, 得到的运行结果的平均值如表 1 所示.

表 1 三种算法聚类结果比较表

性能比较	传统 k-means	基于遗传的 k-means	本文算法
最优解最大/ 最小/均值	143.18/78.94 /97.23	83.56/78.94 /79.86	78.98/78.94 /78.95
平均迭代次数	38	23	15
标准差	0.005	0.002	0.001
聚类准确率	65%	80%	90%

由表 1 的实验结果可知, 传统 k-means 受随机选取的初始聚类中心的影响相当大, 聚类结果不稳定, 多次过早收敛于不同的局部次优解; 遗传 k-means 有时因为染色体的多样性不足而呈现早熟; 本文算法克服了上述缺点, 全局搜索能力显然优于另外两个算法, 聚类结果的稳定性也很强; 同时比较三种算法收敛于最优解的平均迭代次数, 可知本文算法收敛速度最快; 从表 1 还知本文算法聚类准确率明显大大高于另外两个算法, 再有收敛精度也是最优的.

图 1 描绘出三种聚类算法在迭代过程中, 每代种群中的最优染色体的簇内距离和的变化曲线(每种算法 20 次实验的平均值)即目标函数值的收敛曲线.

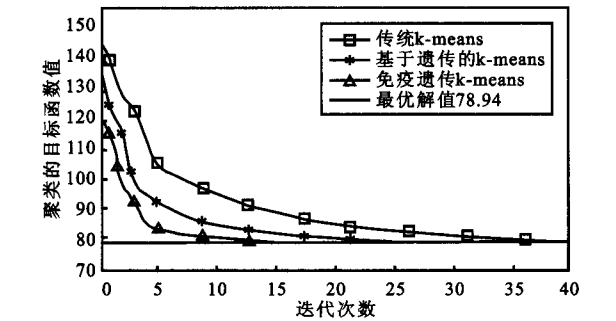


图 1 三个算法的收敛曲线

由图 1 的收敛曲线看出, 基于遗传的 k-means 比传统 k-means 在每次迭代中的目标函数值都要小, 则说明前一算法的聚类效果比后面的算法好. 而本文算法比这两个算法在每次迭代中的目标函数值都要小, 聚类效果是三者中最好的.

5 结束语

本文研究适用于无线传感器网络的分布式 k-means 聚类算法. 针对传统 k-means 算法的缺点, 将免疫机制引入到遗传算法对 k-means 算法进行改进, 提出分布式免疫遗传 k-means 聚类算法. 实验证明了该算法对数据聚类具有明显的优越性能, 收敛速度快, 收敛精度高, 聚类稳定性强准确率也大.

参考文献:

[1] 骆盈盈, 陈川, 毛云芳. 基于传感器网络的 K_均值聚类算法研究[J]. 计算机工程与设计, 2007, 28(6): 1349-1351.

[2] 耿跃, 任军号, 吉沛琦. 基于 K-Means 变异算子的混合遗传算法聚类研究[J]. 计算机工程与应用, 2011, 47(29): 151-153.

[3] 张伟, 廖晓峰, 吴中福. 一种基于遗传算法的聚类新方法[J]. 计算机科学, 2002, 29(6): 114-116.

[4] 行小帅, 潘进, 焦李成. 基于免疫规划的 k-means 聚类算法[J]. 计算机学报, 2003, 26(5): 605-610.

[5] 于璐, 吴振强. 基于免疫的自适应安全模型[J]. 微电子学与计算机, 2009, 26(1): 113-117.

[6] 周洪伟, 原锦辉, 张来顺. 遗传算法"早熟"现象的改进策略[J]. 计算机工程, 2007, 33(19): 201-205.

[7] 唐朝霞. 一种改进的基于遗传算法的 K 均值聚类算法[J]. 成都大学学报: 自然科学版, 2011, 30(2): 162-164.

[8] IrDate Set[EB/OL]. [2012-06-12]. <http://archive.ics.uci.edu/ml/datasets/Iris>.

作者简介:

洪月华 女, (1973—), 硕士, 副教授. 研究方向为数据挖掘、人工智能、无线传感器网络.