

基于改进 C 均值的故障诊断

周德财, 夏士雄, 王志晓

(中国矿业大学 计算机科学与技术学院, 徐州 江苏 221116)

摘 要: 本文提出改进的 C 均值算法, 改进算法引入数据场模型, 通过故障点在数据场中的势值找出噪声点并踢出, 利用故障点在数据场中呈现的自然聚集特性选择初始聚类中心指导 FCM 聚类。仿真实验的结果表明, 改进算法弥补了 FCM 算法的缺陷, 提高了聚类的效率和准确性, 在故障诊断过程中有很好的表现。

关键词: 模糊 C 均值; 数据场; 聚类; 噪声点; 故障诊断

中图分类号: TP301.6

文献标识码: A

文章编号: 1000-7180(2012)11-0120-03

Fault Diagnosis Based on the Improved C-Means

ZHOU De-cai, XIA Shi-xiong, WANG Zhi-xiao

(School of Computer Science & Technology, China University of Mining and Technology, Xuzhou 221116, China)

Abstract: This paper introduce data field model to FCM, and proposed an improved FCM algorithm. On one hand, the improved algorithm uses potential values of fault points in the data field to identify noise points and kicks them out; on the other hand, the improved algorithm utilizes the natural aggregation properties presented by fault points in the data field to select initial clustering centers. The simulation experiment results showed that the improved algorithm has a good performance in fault diagnosis process. The improved algorithm could make up the defects of traditional FCM algorithm, and improve the efficiency and accuracy of clustering.

Key words: fuzzy C-Means; data field; clustering; noise points; fault diagnosis

1 引言

近年来,故障诊断技术^[1]快速发展,研究成果层出不穷。故障诊断主要包括故障检测、数据预处理、故障诊断三方面。故障检测是判断系统中是否出现故障和故障出现的时间点;数据预处理指对采集的数据进行适当的处理,提取故障特征值;故障诊断是指运用故障诊断理论分析推理出故障的类别和严重程度^[2]。现有的故障诊断方法有很多,研究最多的是模糊故障诊断法,其经典的模糊 C 均值方法(FCM)^[3]是一种应用广泛的算法,其基本思想是使得被划分到同一簇的对象之间相似度最大,而不同簇之间相似度最小。由于模糊 C 均值算法能够克服其他故障诊断方法需大量实测样本的限制,其在故障诊断中得到了较多的实际应用。但模糊 C 均值方

法也存在一些缺陷,如:①对孤立点比较敏感;②有可能因随机中心而陷入局部最优;③在修正聚类中心点的过程中,如果初始点选择不当,有可能造成较大的时间消耗。

本文引入数据场模型对模糊 C 均值算法进行改进,提出一种基于改进 C 均值的故障诊断算法(DF-FCM)。该算法利用故障点在数据场呈现的自然聚集特性选择初始聚类中心指导模糊 C 均值聚类,并利用故障点在数据场中的势值找出噪声点并踢出,优化聚类性能。仿真实验的结果表明,改进算法弥补了模糊 C 均值算法的缺陷,提高了聚类的效率和准确性,在故障诊断过程中有很好的表现。

2 数据场

场的概念由英国物理学家法拉第提出,他认为

物体间的非接触相互作用的发生,如万有引力、带电体间的静电力以及磁铁间的磁力等,都必须通过某种中间媒质的传递才能实现,而这种传递相互作用的媒质就是场^[4]. 淦文燕将场论应用于数域空间,提出了数据场的概念,并将数据场理论应用于数据挖掘和知识发现^[5].

数据场理论模拟物理学中场论的思想,认为数据空间中的每个数据都不是孤立的,而是相互影响或作用的,通常把这种相互作用称为能量辐射. 位于数据空间的每个对象都向整个数据空间辐射能量,同时被其他对象辐射,这样就形成一个空间场,称为数据场^[6].

研究发现短程场的势函数能很好地描述数据对象间的相互作用,并证明了具有短程场特点的拟核力场势函数与概率密度函数至多相差一个归一化常数^[4],即拟核力场势函数可以用作总体分布的估计,而最常用的拟核力场是高斯势. 在数域空间中,每个数据点独立向外辐射数据能量,在数域空间中的一点所引起的数据能量之和称为该处数据场的势.

定义 1 已知空间 Ω 中包含数据集 $D = \{x_1, x_2, \dots, x_n\}$ 及其产生的数据场,其中 n 为数据集 D 的个数,令数据对象的位置矢量为 $x_1, x_2, \dots, x_n$,则任一场点 x 处的势值和场强矢量可表示为^[5-6]

$$\phi(x) = \phi_D(x) = \sum_{i=1}^n \phi_i(x) = \sum_{i=1}^n (m_i \times e^{-(\|x-x_i\|/\delta)^2}) \quad (1)$$

$$F(x) = \sum_{i=1}^n ((x_i - x) \cdot m_i \cdot e^{-(\|x-x_i\|/\delta)^2}) \quad (2)$$

其中, $\|x-x_i\|$ 为对象 x_i 到场点 x 的距离; n 为对象数量; m_i 为对象 x_i 的质量,且满足归一化条件,即 $\sum_{i=1}^n m_i = 1$; $\delta \in (0, +\infty)$, 用于控制对象间的相互作用力程,称为影响因子.

影响因子 δ 影响数据场的势分布, δ 越小单个数据点的作用范围越小; δ 越大单个数据点的作用范围越大. 单个数据点的势函数在距离为零处势最大,为 1,在空间无穷远处势最小,趋于 0.

影响因子的取值会影响到最后的结果,如果取太小,所有点彼此孤立,而取太大就会掩盖数据点彼此之间的差距,得不到数据场的关系. 本文采用一种简单试探法,通过选取试探点和比较函数值,使包含极小点的搜索区间不断减小,直至获得满足精度要求的函数极小值. 具体算法 (Optimal_Sigma) 如下:

输入:数据集 $D = \{x_1, x_2, \dots, x_n\}$.

输出:优化的 δ .

① 随机选取 D 中 $n_{\text{sample}} \ll n$ 个样本组成新数据集 D_{sample} ;

② 置 $a = \frac{\sqrt{2}}{3} \min_{i \neq j} \|x_i - x_j\|$, $b = \frac{\sqrt{2}}{3} \max_{i \neq j} \|x_i - x_j\|$, 置精度要求 ϵ ;

③ 令 $\delta_l = a + (1-\tau)(b-a)$, $\delta_r = a + \tau(b-a)$;
 $\tau = \frac{\sqrt{5}-1}{2}$;

④ 计算 $H_l = H(\delta_l)$ 和 $H_r = H(\delta_r)$;

⑤ 当 $|b-a| > \epsilon$ 时,不断执行⑤;

⑥ 如果 $H_l < H_r$, 令 $b = \delta_l$, $\delta_r = \delta_l$, $H_r = H_l$, 计算 $\delta_l = a + (1-\tau)(b-a)$, $H_l = H(\delta_l)$; 否则令 $a = \delta_l$, $\delta_l = \delta_r$, $H_l = H_r$, 计算 $\delta_r = a + \tau(b-a)$, $H_r = H(\delta_r)$;

⑦ 如果 $H_l < H_r$, 则 $\delta = \delta_l$, 否则 $\delta = \delta_r$;

⑧ 返回 δ .

3 改进的 FCM 算法

针对 FCM 算法的不足,本文提出 DF-FCM 算法. DF-FCM 算法的基本思想是将整个故障数据特征空间看作受所有数据点影响的势场,那么故障数据集里的每个数据即为数据点. 如果某数据点位置势值高,说明该数据点周围分布的数据点多;反之,如果某个数据点的势函数值较低,说明其周围邻近的数据点分布稀疏,可以被视为噪声点或野点. 具体过程是计算故障数据点在数据场中的势,踢出噪声点后,构建新的数据场,然后在剩余的数域空间选择极大值点作为初始聚类中心,并将初始聚类中心代入 FCM 进行聚类.

3.1 噪声点检测

噪声点指数据集中与其它数据对象有较大不同的对象,或是那些显著偏离其它数据而能被怀疑为由不同机制产生的对象. 本文假设给定一个阈值(经过多次试验得出),故噪声点可定义为满足势值最小且在给定阈值范围内的数据点. 噪声点检测的具体过程就是计算每个数据点的势,并选择最小势的数据点,如果该数据点满足噪声点条件,将其作为一个噪声点,标记踢出,然后再生成剩余数据点的数据场,再次选择具有最小势值且满足阈值的数据点作为噪声点. 依次类推,直到找出满足条件的所有噪声点. 具体算法如下:

输入:数据集 $X_n = (x_1, x_2, \dots, x_n)$, 阈值 ϵ .

输出:噪声点集合.

① 计算数据集中两两数据向量之间的欧式距离 d , 存入距离矩阵;

② 根据式(1)计算数据场内各数据点势大小,找出具有最小势值且达到阈值的数据点,踢出并存入噪声点集合;

③ 重复①步骤直到满足要求.

3.2 基于数据场的初始簇中心确定

考虑到势场分布的局部极大值点就相当于数据场的“虚拟场源”,所有数据对象在各自的“虚拟场源下”自组织聚集特性,因此可以将势场分布的局部极大值视为聚类中心,形成数据的初始划分.在踢出噪声点之后,对剩下的非噪声点数据重新构建数据场,通过采用线性收敛的区间分析法搜索数据场的所有临界点,然后根据 Hesse 矩阵的特征值确定局部极大值,作为数据集的初始聚类中心.

输入:数据域空间 X .

输出:极大值点集合.

① 构建消除噪声点后的数据场;

② 搜索数据场分布的所有拓扑临界点;

③ 根据 Hesse 矩阵的特征值确定局部极大值.

3.3 DF-FCM 算法描述

DF-FCM 算法涵盖影响因子优化、噪声点检测、确定初始聚类中心和 FCM,最终得出聚类结果,算法如下:

输入:数据集 $D = \{x_1, x_2, \dots, x_n\}$.

输出:数据的聚类结果.

① 影响因子优化;

② 检测噪声点;

③ 确定初始簇中心;

④ 将③返回结果带入 FCM 算法,不断优化;

⑤ 返回聚类结果.

算法分析

DF-FCM 算法过程开销最大的是计算所有点的距离,时间复杂度为 $O(n(n-1))$,总体时间复杂度是 $O(n^2)$,而 FCM 算法时间复杂度也是 $O(n^2)$,所以 DF-FCM 算法与 FCM 算法时间复杂度基本维持在一个层次.

4 仿真实验

为验证算法有效性,本文分别选取 Aggregation 数据集和真实电机故障实验数据,实现模糊 C 均值算法和改进的 C 均值算法作为比较.所有程序用 python 实现,实验环境为:处理器 2.94 GHz,内存 3 GB,硬盘 320 GB,操作系统 Windows 7.

综合实验结果,可以看出使用改进的 FCM 算法在寻找故障中心的速度和故障诊断的准确率优于 FCM 算法.

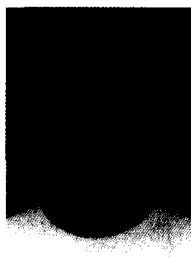
5 结束语

由于模糊 C 均值算法的缺陷突出,针对其聚类结果对初始中心敏感的特性,结合数据场的思想,在踢出故障数据噪声点后给出初始聚类中心指导 FCM 进行聚类,然后运用于故障诊断,提高了聚类精度,减少了整体迭代次数.但是,在簇边缘点模糊归类上还有不足,因此下一步的研究工作主要是提高簇边缘点模糊归类精度.

参考文献:

- [1] Zhixiong Li, Xinping Yan, Zhiwei Guo. A new intelligent fusion method of multi-dimensional sensors and its application to tribo-system fault diagnosis of marine diesel engines[J]. Tribol Lett, 2012(47):1-15.
- [2] 何勇,李增芳. 智能化故障诊断技术的研究与应用[J]. 浙江大学学报:农业与生命科学版, 2003, 2(2): 119-124.
- [3] Zhang Yufang, Luo Junwei, Xiong Zhongyang. Study on class merging cluster algorithm based on fuzzy c-means[J]. Computer Engineering and Applications, 2009, 45(24): 122-124.
- [4] 李德毅,杜鹁. 不确定人工智能[M]. 北京:国防工业出版社, 2005.
- [5] 涂文燕. 一种基于数据场的层次聚类方法[J]. 电子学报, 2006, 34(2): 258-262.
- [6] 王莉军. 一种基于数据场的社区发现算法[J]. 计算机应用研究, 2011(11): 4142-4145.
- [7] Gionis A, Mannila H, Tsaparas P. Clustering aggregation[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2007, 1(1): 1-30.
- [8] 刘曼兰. 永磁直流电机故障在线监测与智能诊断的研究[D]. 哈尔滨:哈尔滨工业大学, 2007.

作者简介:



周德财 男, (1988—), 硕士研究生. 研究方向为数据挖掘.

夏士雄 男, (1962—), 教授, 博士生导师. 研究方向为矿山数字化和智能处理.

王志晓 男, (1979—), 副教授. 研究方向为数据挖掘.