

一种拟人聚类算法在 PHM 聚类分析中的应用

贺呈磊, 唐 磊, 刘 曦

(西安微电子技术研究所, 陕西 西安 710054)

摘 要: 提出了一种基于数据密集度和邻域的拟人聚类算法. 该算法模仿人在观察分类时的思维过程, 通过约束性的全局搜索和对初始聚类中心的优化合并, 最终得到规定数目的聚类中心从而完成整个聚类过程. 利用 PHM 硬件模拟平台的数据进行验证, 该聚类算法较 FCM 算法在局部的二次聚类的准确性上有约 5 倍的提高.

关键词: 健康管理; 聚类; 模糊 C 均值算法; 拟人; 聚类中心

中图分类号: TP391

文献标识码: A

文章编号: 1000—7180(2016)09—0032—04

Application of an Anthropathic Clustering Algorithm to PHM

HE Cheng-lei, TANG Lei, LIU Xi

(Xi'an Microelectronics Technology Institute, Xi'an 710054, China)

Abstract: An anthropathic clustering algorithm based on the data set's intensity and neighborhood was proposed in this paper. The algorithm imitates human's thinking procedure of classification, and it obtains the given number of centers and accomplishes the clustering by searching the whole data set globally and restrictedly, incorporating and optimizing the initial centers. The experiment using the data from the hardware emulational platform of PHM indicates that the algorithm has the accuracy on local and secondary clustering about 5 times larger than Fuzzy C Means algorithm.

Key words: PHM; clustering; fuzzy C means algorithm; anthropathy; cluster centers

1 引言

作为未来自主保障系统的重要基础, 预测和健康管理(PHM)关键技术的研究和应用在提高系统的安全性、经济性方面的巨大潜力受到了人们的广泛关注, 世界各国在航空、航天等复杂领域的系统研制过程中都开展了健康管理相关技术的研究^[1]. 在进行自主健康管理研究时, 首要问题是利用能够表征系统状态的“健康信息”参数, 建立系统在不同工作状态下的标准, 以便在此基准上对系统运行时的状态进行诊断、评估和给出相应的决策建议^[2]. 为建立相应的标准, PHM 需要对系统的历史数据进行聚类分析以区分不同的工作状态, 由此, 聚类分析既是整个健康管理的基础, 也是健康管理技术的研究方向之一.

本文在进行空间电子系统健康管理的课题研究

时, 以系统的不同工作模式为依据进行分类^[3], 在 PHM 的聚类分析中应用 FCM 算法时, 由于系统不同状态下数据的特点, 其在进行空闲模式和正常模式的局部第二次分类时未能进行正确区分, 这是因为 FCM 算法在迭代时陷入了局部极值点. 针对 FCM 得不到全局最优解的问题, 国内外许多学者进行了相关的研究^[4-7], 但并不完全适用于 PHM 的聚类分析.

针对 PHM 聚类分析时应用 FCM 算法出现局部分类错误的问题, 本文提出了一种基于密度和邻域的快速划分式算法.

2 FCM 算法在 PHM 中的应用

模糊 C 均值聚类算法是一种基于划分的聚类算法, 它是在“硬”C 均值聚类算法(HCM)的基础上加入相应的模糊元素而产生^[8], 其核心思想是得到

样本集中每个样本点对 C 个聚类中心的隶属度,再依据隶属度最大的原则将每个样本点划分到相应的类别中。

2.1 FCM 算法的原理和分析

设将样本集 $X = \{x_1, x_2, \dots, x_n\}$ 划分成 c 类 $X_1, X_2, \dots, X_C$, FCM 算法的步骤为:

步骤一 随机选择 c 个初始聚类中心点 $c_i, i = 1, 2, \dots, c$;

步骤二 计算所有样本与各聚类中心的欧式距离 $d_{ij}, i = 1, 2, \dots, c, j = 1, 2, \dots, n$;

步骤三 更新隶属度矩阵 U , 矩阵元素为 $u_{ij}, i = 1, 2, \dots, c, j = 1, 2, \dots, n$;

步骤四 计算目标函数 J 的函数值, $J(U, c_1, \dots, c_c) = \sum_{i=1}^c J_i = \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m d_{ij}^2$, 其中 m 为加权指数;

步骤五 设定一个极小值 ξ , 计算相邻两次目标函数的差值 η , 若 $\eta < \xi$, 则算法停止, 否则, 更新聚类中心 c_i , 转步骤二。

由上述算法的描述看出, FCM 算法的目标函数值为每个样本到各聚类中心的距离加权的平方和, 算法的迭代过程就是使目标函数值最小化的过程。由于目标函数可能有多个极值点, 初始聚类中心的选取是否在极值点附近将严重影响该算法的性能, 因此随机选取的初始聚类中心并不能保证 FCM 算法能够收敛到最优解, 也即 FCM 算法有时会得不到正确的聚类结果, 尤其是在样本集中各类样本数目差距较大的情况下^[3]; 若利用经验人为地选取聚类中心则失去了课题研究的无监督性和自主性, 而且这在样本集中样本数目较大时也将影响到应用的可行性。

2.2 FCM 算法在 PHM 聚类分析中的应用

PHM 进行聚类分析的数据集来源于系统运行时的状态, 例如某电子系统运行时, PHM 在每一采样时刻得到系统的电压值 V 、电流值 I 及温度值 T 等信息组成一个三维的数据样本点, 而一段时间内得到的多个样本点则组成了原始的数据样本集。

图 1 中所使用的数据集如第 3 节实验中所述, i 为全局数据集, 其中左下角为休眠模式, 右上角为故障模式, 中间为空闲、正常及其过渡状态, iii 为中间局部数据集的放大情形, 而 ii 和 iv 为聚类结果, 其中的“ x ”为 i 和 iii 对应的聚类中心。应用 FCM 时, 算法可以将故障模式和非故障模式区分开来, 但没有将非故障模式中的空闲模式、正常工作模式及二者之间的过渡状态区分开来, 如图 1 中的 i 和 ii 所

示; 而在应用 FCM 算法对局部的非故障模式进行第二次分类时依然出错, 如图 1 中的 iii 和 iv 所示。如上一小节中算法分析所述, 图中 iii 在应用 FCM 算法时由于局部样本集中各类不同的样本点数差距较大、样本点比较集中及界限不明晰而导致聚类结果出错。

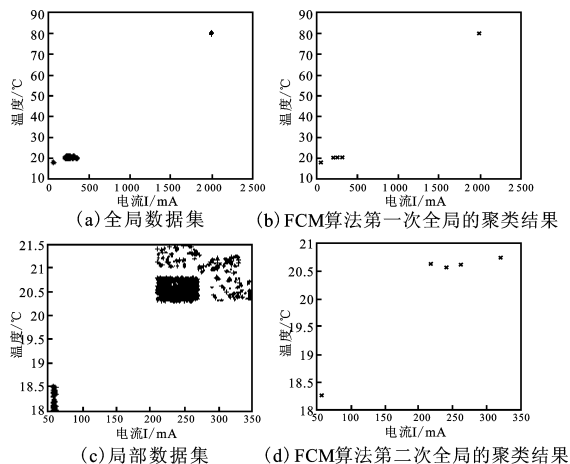


图 1 FCM 算法在 PHM 中的应用

3 拟人聚类算法

3.1 拟人聚类算法的思想

一般情况下, 若给定一个二维平面上的点集而要求一个人在进行分类时, 其通常会首先进行观察并找出点集中比较密集的多个区域作为分类的基础, 这些区域都具备样本点数少、范围粗略且相互之间距离较大等特点; 然后在这些多个基础分类的前提下, 分别寻找各自周围最近的点, 同时按照类内相互间距最小, 类间相互间距最大的原则进行判断, 将符合要求的点划分到对应的基础分类中从而形成新的基础分类, 同时新基础分类的中心点也发生了变化, 即动态的分类; 依照上述思路反复进行下去, 最终完成整个平面点集的划分。上述的观察聚类过程示意如图 2 所示。

上文中提到的类内、类间的距离度量是通过设定相应的阈值体现出来的。考虑一维情形下, 若需要在坐标轴上进行均匀分割时, 首先要找出轴上距离最远的两个端点, 再将最大距离按照分段的数目进行均匀分割。由此, 相邻段中心的之间距离即为上述的均匀分割值, 而每个分段的段内半径大小即为该均值的一半。若将其推广到二维或更高维的情形中并且考虑最坏的情形, 可以将距离适当放大, 即参考一维情形中的距离分割。

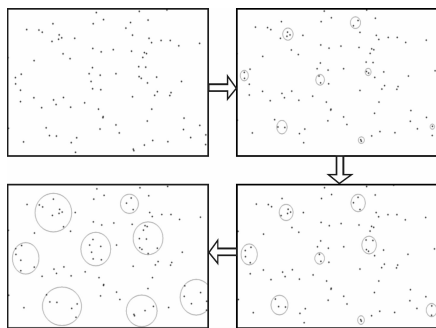


图2 观察聚类过程的示意图

3.2 拟人聚类算法的基本原理

本文仿照上述的分类思想提出了一种新的聚类算法,该算法以样本点之间的距离作为密集度的选择依据,通过自适应计算邻域中心的间距阈值和简单的合并优化快速得到所需的聚类中心.具体方法为:首先遵循密度最大的原则,选择未分类的数据集中距离最近的两个点并取其均值作为一个邻域的中心;然后按照就近原则,并结合邻域内的点和邻域中心之间的距离必须小于设定阈值、邻域中心之间的距离必须大于另一设定阈值的原则,计算出数据集中满足条件的一个点,将其纳入该邻域同时更新邻域中心和数据集,如此动态地进行循环迭代划分,迭代结束时的邻域中心就是一个预取聚类中心;多次进行上述搜索直到找到所有满足条件的预取聚类中心,最后按照分类数目和距离合并而得到需要的 C 个聚类中心.其中上述阈值由点集中各点之间的距离的最大值及聚类数目 C 决定;由于执行的是全局性搜索,所得到预取聚类中心个数一般都大于预设的聚类数目 C ,因此本算法在此基础上加入了中心点合并优化的过程.

设欲将样本集 $X = \{x_1, x_2, \dots, x_N\}$ 分成 c 个类,则该算法的基本步骤为:

步骤一 初始化:计算并找出样本集 X 中所有样本点之间的最大距离 $D_{\max}$,计算邻域中心之间的距离阈值 $R = \frac{D_{\max}}{c}$,邻域内的点到邻域中心之间的距离阈值 $r = \frac{R}{2}$;

步骤二 初始邻域及邻域中心:计算并将当前数据集中相互之间距离最近且均满足设定的类间阈值条件 R 的两个数据点 $x_i, x_j, 1 \leq i, j \leq n$ 作为初始邻域,均值点 $\bar{x}_k$ 作为相应初始邻域的中心;

步骤三 预取聚类中心:计算并将当前数据集中距离邻域中心最近且满足设定的类内阈值条件 r 和

类间阈值条件 R 的数据点 $x_i, 1 \leq i, j \leq n$ 划分到当前邻域中,删除该数据点后更新当前的数据集 X 及其邻域中心 $\bar{x}_k$,反复执行步骤三,并将最后的邻域中心 $\bar{x}_k$ 作为预取聚类中心 c_k ;

步骤四 k 个预取聚类中心:迭代执行步骤二和步骤三,得到所有 $k, k \geq c$ 个满足条件的预取聚类中心 c_k ;

步骤五 c 个聚类中心:若预取聚类中心个数 $k > c$,则将 k 个初始聚类中心中距离最近的两个进行合并同时更新预取聚类中心个数 k ,反复执行步骤五,并将最后的 c 个预取聚类中心作为最终的 c 个聚类中心.

其中算法的步骤二和步骤三所执行的阈值 R 和 r 均来自于步骤一,其设定是按照类间间距最大和数据分布的最坏情形进行的.该设定使算法在进行每一个初始邻域中心的选取和对数据集的动态划分时,能够始终满足类内及类间的“安全”间距,而非中心点固定的划分.本算法在全局性的搜索结束时所得到的预取聚类中心数目 k 一般情况下不小于所要求的聚类数目 c ,考虑到算法实现时对复杂度、耗时以及精确度的应用要求,算法在步骤五中简单地以两个预取聚类中心所包含的类内元素个数为权重进行合并处理,从而快速得出 c 个聚类中心.本聚类算法的流程图如图 3 所示.

4 实验结果

为了模拟电子系统的各类非故障和故障情形,同时验证本文提出的算法,本次实验利用课题设计和搭建的 PHM 硬件平台,使用其中的故障模拟板所采集到的数据进行验证.实验时使用包括故障模拟板采集到的 30 组数据及 MATLAB 按照边界阈值补充的共计 660 组数据的数据集进行评估.

如图 4 所示,实验时采集到的每组数据具有 3 个属性值:电流 I 、电压 V 和温度 T ,聚类结束后的聚类中心具有相同的属性值.考虑到实际应用中,非故障模式下所采集到的电压值 V 的基本不变,本实验只使用电流和温度这两个属性值所生成二维的数据集,其在图中分别作为 x 轴和 y 轴.图 4 使用 FCM 算法和本文的拟人算法的某次实验结果.

图中的数据值是为方便模拟故障情形而由故障模拟板设定出的值.图 4 中,ii 和 iv 中的“ x ”为 i 和 iii 数据集中相应的类中心,其中左下角为系统处于休眠模式下的情形;中间数据点最多的部分为系统处于空闲状态下的情形;其他 3 类为系统正常工作

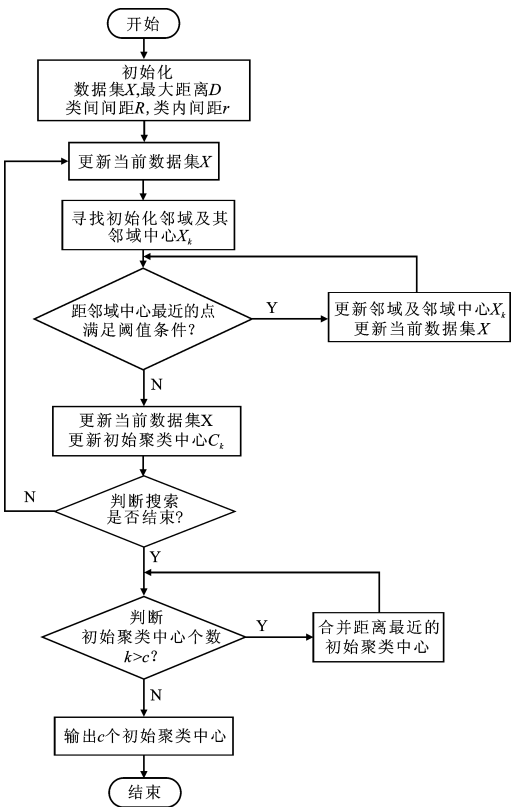


图 3 新聚类算法的流程图

文提出的拟人算法的聚类结果,其结果是正确的.此外,本次实验中 FCM 算法和拟人算法所用的时间分别为 0.04 s 和 0.06 s.

为更好的验证本算法的有效性,本次实验在 MATLAB 中按照相同的方法进行了 100 次,统计结果如表 1 所示.

表 1 100 次试验结果对比统计

算法	正确次数	平均耗时/s
FCM 算法	12	0.043 5
拟人聚类算法	71	0.056 5

由表 1 的统计结果可以看出,本文提出的聚类算法的正确次数要明显优于 FCM 算法,虽然新算法相对于 FCM 算法在耗时方面有所增加,但由于本算法都是基于局部邻域的距离运算,并没有 FCM 算法中的全局性的大量的矩阵运算,同时考虑到空间电子系统 PHM 应用时预设的采集周期为分钟级甚至更高,而且是在经过一定时间长度的采集之后才运用该算法对多组数据进行处理,因此折中考虑来说,新聚类算法更需要注重其性能上的提升,运算耗时的增大在一定程度上是可接受的,本文的改进算法在局部的分类应用上是合理的.

5 结束语

空间电子系统 PHM 的局部数据集存在着类别之间界限不明晰且相互之间数目相差较大的特点,应用 FCM 算法进行局部的第二次聚类时极易得到错误的分类结果,而相应的改进算法并不适用于本次应用.为此,本文仿照人在观察分类时的思维过程,提出了一种基于密集度的简单快速的聚类算法^[9],本算法自适应地设定聚类过程中的类内和类间的距离阈值,同时结合类内间距最小和类间间距最大的分类原则进行全局性的搜索得到合理的聚类中心.经实验结果验证,该算法较 FCM 算法明显提高聚类的准确性,具有较好的可行性和有效性.

参考文献:

[1] 罗荣蒸,孙波,张雷,等. 航天器预测与健康管理技术研究[J]. 航天器工程,2013,22(4): 95-102.
[2] Nikhil M Vichare, Michael G Pecht. Prognostics and health management of electronics[J]. IEEE Transactions on Components and Packaging Technologies, 2006,29(1): 222-229.

(下转第 41 页)

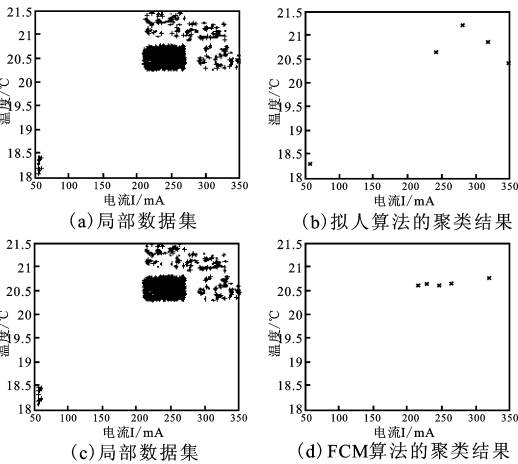


图 4 实验结果对比

模式以及模式之间的过渡情形,即右侧为系统从空闲模式切换到正常模式时电流突然增大而温度缓慢变化的过程,右上侧为系统在正常模式下经过一段时间后稳定的情形,上侧为系统从正常模式切换到空闲模式时电流突然变小而温度缓慢变化的过程,三者按照 1:50:5:5:5 的比例模拟系统实际运行时各类别的数目.如上的实验结果中,iv 为 FCM 算法的局部二次聚类结果,其迭代计算时陷入了局部极值,因而聚类结束后给出了错误的聚类结果,ii 为本

一站式跨平台校园微门户建设,提出使用 RSA 与 MD5 联合应用的敏感数据保护机制,保障敏感数据安全;同时利用微信公众平台进行二次开发建设校园门户,极大地拓展了传统校园信息门户和校园移动门户在移动互联网领域中的覆盖率;且可在多类型智能移动操作系统运行。

本文的下一步研究计划是充分利用实名制校园微门户的客户端反馈及上下文信息,完成针对用户真实身份的大数据分析与个性化服务推荐,从而逐步实现校园微门户及核心业务的“互联网+”智慧化,为校园信息化的十三五规划做好支撑工作,更好的提升校园信息化水平。

参考文献:

[1] 童华章,吴颖骏,方伟杰,等. 学生一站式服务中心的研究与设计[J]. 中国海洋大学学报:自然科学版, 2008, 49(Sup1):33-37.

[2] 郑锋. 一站式服务架构下的电子政务应用系统[J]. 计算机工程, 2007, 33(14):249-250.

[3] Park J H, Chao H C, Chen W E, et al. Special Issue on information security and data protection in future generation communication and networking[J]. Wireless Personal Communications, 2009, 51(1):1-3.

[4] Fernandes D A B, Soares L F B, Gomes J V, et al. Security issues in cloud environments: a survey[J]. In-

ternational Journal of Information Security, 2014, 13 (2):113-170.

[5] Zhang H P, Zhang R Q, Zhao Y P, et al. Big data modeling and analysis of microblog ecosystem[J]. International Journal of Automation & Computing, 2014, 2 (2):119-127.

[6] Tsai C W, Lai C F, Chao H C, et al. Big data analytics: a survey[J]. Journal of Big Data, 2015, 2(1): 1-32.

[7] 席新亮. 微信公众平台 JSSDK 开发实战[M]. 北京:电子工业出版社,2015:1-2.

[8] 腾讯公司. 微信公众平台开发者文档[EB/OL]. [2015-12-09]. <http://mp.weixin.qq.com/wiki/11/434109e8de46b3968639217bbcb16c2f.html>.

作者简介:

洪丹丹 女,(1981-),硕士,工程师.研究方向为移动互联网技术、校园信息化关键技术.

罗军锋 男,(1976-),硕士,高级工程师.研究方向为大数据挖掘、高校信息化.

冯兴利(通讯作者) 女,(1986-),硕士,工程师.研究方向为校园信息化. E-mail:fengxingli 5927@sina.com.

徐 墨 女,(1978-),硕士.研究方向为校园信息化.

锁志海 男,(1971-),博士研究生.研究方向为数据库应用技术、高校教育信息化.

(上接第 35 页)

[3] 刘蕊洁,张金波,刘锐. 模糊 c 均值聚类算法[J]. 重庆工学院学报(自然科学),2008,22(2): 139-141.

[4] 张红旗,王春光,李海军. 基于遗传算法的草莓图像 FCM 分割方法研究[J]. 农机化研究,2015, 36(4): 55-58.

[5] 苏锦旗,张文字,薛昱. 基于改进人工免疫方法的混合模糊聚类算法[J]. 计算机系统应用,2014, 23(8): 194-197.

[6] 张慧哲,王坚. 基于初始聚类中心选取的改进 FCM 聚类算法[J]. 计算机科学,2009,36(6):206-209.

[7] 陈兴蜀,吴小松,王文贤,等. 基于特征关联度的 K-means 初始聚类中心优化算法[J]. 四川大学学报(工程科学版),2015,47(1): 16-19.

[8] 杨纶标,高英仪,凌卫新. 模糊数学原理及应用[M]. 5 版:广州:华南理工大学出版社,2011. 7.

[9] 谢娟英,高 瑞. Num-近邻方差优化的 K-medoids 聚类算法[J]. 计算机应用研究,2015,32(1):30-34.

作者简介:

贺呈磊 男,(1989-),硕士研究生.研究方向为预测与健康管理技术. E-mail:he_303560@163.com.

唐 磊 男,(1973-),研究员,硕士生导师.研究方向为空间嵌入式计算机系统结构及大容量固态存储技术.

刘 曦 男,(1979-),研究员.研究方向为卫星综合电子系统技术.