文章编号:1006-2467(2006)03-0439-05

一种基于群体智能聚类的设备性能横向比较算法

张 蕾[1], 曹其新[1], 李 杰[2]
(1. 上海交通大学 机器人研究所,上海 200030;
2. 威斯康星大学 智能维护中心,密尔沃基 WI 53224)

摘 要: 提出了一种基于群体智能的设备性能横向比较算法,该算法将设备模式投影于二维平面上,然后依据群体智能聚类,实现设备性能的自组织聚类分析. 为了提高群体智能聚类算法的运行效率,提出了利用主成分分析改善模式投影时的随机性. 此外,为了减小参数选取对算法的影响,提出了一种更简单的相似度衡量公式. 将改进后的算法与原算法进行比较,结果显示,改进算法的运行效率更高. 利用改进算法对某中央空调的数据进行分析,验证了算法能满足设备性能自组织聚类的要求.

关键词: 智能维护; 聚类算法; 主成分分析
中图分类号: TP 18 **文献标识码**: A

An Algorithm for Comparing Machine Performance Based on Swarm Intelligence Clustering

ZHANG Lei[1], *CAO Qi-xin*[1], *LEE Jay*[2]
(1. Research Inst. of Robotics, Shanghai Jiaotong Univ., Shanghai 200030, China;
2. Research Center of Intelligent Maintenance Systems,
The Univ. of Wisconsin, Milwaukee WI 53224, USA)

Abstract: An algorithm for assessing the machine performance based on swarm intelligence was proposed. The patterns representing the machine performance are first mapped into a plane, then analyzed by the clustering algorithm based on the swarm intelligence. To improve the running efficiency of the algorithm, the stochastic mapping of the patterns was modified based on principal component analysis. Moreover, to reduce the influence on the algorithm caused by the parameters selection, a simple way to calculate the similarities of the patterns was presented. Finally, the modified algorithm was compared with the original one and the results prove that the efficiency is improved significantly. The modified algorithm is applied in analyzing the data of air-condition which proves it can cluster the data with self-organizing characters.

Key words: intelligence maintenance; clustering algorithm; principal component analysis

随着 Internet 和 Web 技术的飞速发展,基于网络的智能维护技术成为设备维护领域的一种新的发展趋势. 与传统的反应式维护不同,智能维护的基本理念是主动式的预测和预防维护,如何准确、快速地

收稿日期:2005-04-14
基金项目:国家自然科学基金资助项目(50128504);上海市科委重大攻关资助项目(04DZ12006)
作者简介:张 蕾(1975-),女,山东潍坊人,博士生,主要从事远程监控与智能维护方面的研究.
曹其新(联系人),男,教授,博士生导师,电话(Tel.)021-62932750.

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

对设备性能的退化进行评估是实现智能维护的关键.一般设备的性能评估有纵向和横向比较.前者是根据设备自身的历史数据评估性能的变化趋势;后者是与工作时间相同的同类产品进行性能比较,评估自身的性能状态.对于纵向比较算法:Lee 等[1,2]最先提出了一种基于小脑模型神经网络(CMAC)的方法,并在步进电动机和机器人上进行了验证;Lin 等[3]提出了一种基于增强 CMAC 的旋转机械性能退化分析方法;Yan 等[4]用逻辑回归法来评估电梯门系统的性能退化;Dae 等[5]提出了一种基于自适应谐振理论的方法,提高了算法对设备不同运行状态的适应能力.对于横向比较算法,目前的研究还处于起步阶段.然而从单台设备的性能诊断转向多台设备性能的对比是智能维护改变传统维护的一个重要方面,也是网络环境优势的充分体现,因此,开发设备性能的横向比较算法非常必要.

群体智能是指简单个体通过交互和协作而表现出来的一种复杂智能行为,设备性能的横向比较就是要通过"设备群体"之间的相互作用而得到评估设备性能状况这一"智能"行为.根据这一结合点,本文提出利用群体智能聚类(Swarm Intelligence Clustering, SIC)算法将各设备按照性能状况聚类,从而实现设备性能的横向评估.

1 基于群体智能的聚类算法

某些种类的蚂蚁能够将分散在穴内的蚂蚁尸体分拣成堆,Deneubourg 等[6]提出了解释这种行为的基本模型,称为 BM 模型.Lumer 等[7]将 BM 模型推广应用到数据的聚类分析,主要思想是将待聚类模式随机分布在一个二维网格上,然后由蚂蚁测量当前对象在局部环境内的群体相似度,并将这种群体相似度通过概率转换函数转换成拾起或放下的概率,通过群体之间的这种相互作用,经多个循环后即可实现模式的聚类.

群体相似度是一个待聚类模式(对象)与其所在一定的局部环境中所有其他模式的综合相似度[8].群体相似度的基本计算公式为

$$f(O_i) = \sum_{O_j \in \mathrm{Neigh}(r)} \left[1 - \frac{d(O_i, O_j)}{\alpha} \right] \tag{1}$$

式中:Neigh(r) 为模式 O_i 的邻域,是以 r 为半径的圆形区域,也称为 O_i 的局部环境;$d(O_i, O_j)$ 为模式 O_i 与 O_j 之间的距离,通常为欧式距离;α 为群体相似系数,它的大小对聚类中心的个数以及算法的收敛速度具有重要的影响.

概率转换函数是将群体相似度转换为简单个体(蚂蚁)移动待聚类模式概率的函数.文献[8]中采用的是一种比基本模型简单的概率转换函数,它是 2 条斜率为 k 的直线,分别对应拾起和放下概率.

$$P_p = \begin{cases} 1 - \varepsilon & f(O_i) \le 0 \\ 1 - k f(O_i) & 0 < f(O_i) < 1/k \\ 0 + \varepsilon & f(O_i) \ge 1/k \end{cases} \tag{2}$$

$$P_d = \begin{cases} 1 - \varepsilon & f(O_i) \ge 1/k \\ k f(O_i) & 0 < f(O_i) < 1/k \\ 0 + \varepsilon & f(O_i) \le 0 \end{cases} \tag{3}$$

式中,ε 为一个很小的实数.设定一个阈值 P_r,通过比较 P_p、P_d 与 P_r 的大小来决定模式是否拾起或放下.经过上述简化后,概率转换函数的参数只有 k,而基本模型中,概率转换函数的参数包括 k_1 和 k_2 值,因此简化了参数选取的复杂度.但是,在具体应用时,k 应该怎样合理设置,缺乏相应的理论指导.群体智能聚类算法的实现步骤可参见文献[8,9].算法的结束条件可以有两种:一是达到最大循环次数;二是平面内没有模式再被移动.由于后者计算比较复杂,一般都采取第一种.

2 改进的 SIC 算法

针对 SIC 算法的效率不高和参数选取复杂两个问题,本文提出一种改进的 SIC 算法(Modified Swarm Intelligence Clustering, MSIC).

2.1 利用主成分分析改善初始坐标投影的随机性

在 SIC 算法中,初始模式被随机地放在一个二维网格的平面中,每个模式对应一个坐标,由于坐标是随机确定的,因此,在聚类的初始阶段,每个模式与其局部环境中的其他模式的相似度比较低,这导致模式很容易被拾起,却不容易被放下.

怎样使得在 R^n 维空间比较相近的模式投影到二维空间后也比较相近,本文提出首先对各模式进行主成分分析,保留前 2 个主成分,并且对这 2 个主成分进行一定的处理,如取整、放大、平移,处理的目的是使各主成分都为非负整数而且容易区分.根据主成分分析的原理,前 2 个主成分可保留数据的大部分信息,各模式以对应的 2 个主成分的处理值作为坐标进行投影,保证了在 R^n 维空间比较相近的模式,投影到二维空间后也比较相近.这样在聚类初始时刻,各模式投影后,就和周围局部环境内的模式具有很大的相似性,相当于原 SIC 算法已经执行了多个循环后的结果,因此算法的运行次数会大大缩减.

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2.2　改善的相似度计算公式

在 SIC 算法中,数据的相似度根据式(1)计算,其中相似度系数 α 直接影响聚类中心的个数,同时也影响算法的收敛速度. 若 α 过大,则不相似的对象可能会聚为一类;若群体相似系数过小,则相似的对象可能会聚到不同的类中. 文献[8]中指出,α 应随着循环次数的增加逐渐变化,但在算法的具体实现中,α 该怎样变化缺乏相应的理论指导,不同的应用中 α 的变化不尽相同,因此难以掌握. 另外,其概率转换函数采取式(2)和(3)中的直线方式,虽然起到了简化参数选取的作用,但可以看到,k 值对拾起和放下概率也会产生重要影响,因此,怎样合理选择 k 值也是一个问题.

为了避免 α 和 k 这些参数取值对聚类结果的影响,本文提出更简单的相似度衡量方式:由式(1)可知,相似度衡量的本质由模式之间的距离决定,因此直接采用模式之间的距离来表示相似度. 模式 O_i 与其局部环境中其他模式的相似度为

$$f(O_i) = \frac{1}{n}\sum_{O_j \in Neigh(r)} d(O_i, O_j) \tag{4}$$

式中:n 为模式 O_i 的局部环境内其他模式的个数,其他符号的含义同式(1). 这种相似度计算方法物理意义很明显,$f(O_i)$ 的值越大,说明模式 O_i 的群体相似度越小.

与 SIC 算法不同,改进的算法不再采用概率转换函数,而是设定相似度阈值 F,$f(O_i)$ 与阈值进行比较决定模式是被拾起还是被放下. 这种计算方法简单易行,而且避免了 k 的取值对算法的影响. 由于在聚类的初始阶段,模式之间的距离比较大,F 应取较大的值. 随着循环次数的增加,相似的模式之间慢慢聚在一起,$f(O_i)$ 会逐渐减小,这时应调整阈值 F,使其也逐渐减小. 本算法中,F 的调节公式为

$$F(t+1) = \begin{cases} F(t) & \mathrm{mod}(t,500) \neq 0 \\ kF(t) & \text{其他} \end{cases} \tag{5}$$

式中,k 可取 0.7～1.0 的实数. 即每 500 次循环后,$F(t)$ 的值就相应减少一点,t 为循环次数. 当然,每 500 次也不是绝对的,可根据相似度下降的快慢适当加大或缩小.

2.3　MSIC 和 SIC 算法比较

MSIC 算法如下:

输入:待聚类模式.

输出:聚类的各模式.

(1) 初始化参数:cycle_number(最大循环次数),ant_number(蚂蚁数量),半径 r,相似度阀值初始值 $F(1)$,阈值调整系数 k.

(2) 对待聚类模式进行主成分分析,保留前 2 个主成分,并对其进行一系列处理,作为每个模式投影到二维网格平面上的坐标.

(3) 给一组蚂蚁赋初始模式值,并且把模式的坐标赋给蚂蚁,蚂蚁初始状态为无负载.

(4) for $i = 1$:cycle_number

　for $j = 1$:ant_number

　　以蚂蚁 j 的初始模式为中心,r 为半径,利用式(4)计算此模式在局部环境中的相似度 f.

　　若 load_ant(j) = 0,比较 f 与阈值 $F(i)$ 的大小,如果 $f \leq F(i)$,则蚂蚁拾起此模式,蚂蚁负载 load_ant(j) = 1,随机再赋给蚂蚁一个坐标值;否则,蚂蚁不拾起此模式,随机再赋给蚂蚁新的模式值和坐标值.

　　若 load_ant(j) = 1,如果 $f > F(i)$,蚂蚁放下此模式并将蚂蚁当前的坐标赋给该模式,load_ant(j) = 0,随机再赋给蚂蚁一个新的模式和坐标值;否则,蚂蚁不放下此模式,随机再赋给蚂蚁新的坐标值.

　End (蚂蚁个数循环结束)

　按照式(5)计算相似度阈值 $F(i)$

End(达到最大循环次数)

(5) 根据聚类结果计算每类的聚类中心,输出各聚类的模式.

本文利用 MSIC 算法与文献[8]中的 SIC 算法进行了比较分析,试验数据来自文献[10]中的 UCI 机器学习数据库,选用鸢草(iris)数据集. 共有 150 条记录,每条 4 个属性,类别数 3. 两种算法的参数如表 1 所示. 由表可见,MSIC 算法避免了 α 和 k 值的选取不当对算法的影响.

表 1　两种算法的参数比较

Tab. 1　Comparison of the two algorithms

SIC 算法	MSIC 算法
最大循环次数 60 000	最大循环次数 10 000
蚂蚁个数 6	蚂蚁个数 6
投影平面 480 ×450	投影平面 80 ×80
半径 $r = 20$	半径 $r = 5$
$\alpha = 0.3$～0.4	-
$k = 0.1$ (式(2)、(3)中)	$k = 0.95$ (式 5 中)
P_r(未说明)	$F(1) = 1.8$

MSIC 算法中,各模式首先进行主成分分析,保留前 2 个主成分. 经计算可知,前 2 个主成分包含了原数据量 97.76 %的信息. 对主成分值做如下处理:

(1) 放大. 由于主成分的数值较小,为便于比较,先将各数均乘以 10.

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

(2) 取整.

(3) 平移. 求出第 1 和第 2 主成分的最小值,每对主成分分别减去对应的最小值,得最后的处理值.

处理的目的是使得投影坐标落在第一象限内. 处理后的 2 个主成分值作为相应模式的横纵坐标值进行投影(第 1 主成分对应横坐标,第 2 主成分对应纵坐标).

两种算法中各模式初始时刻的投影如图 1 所示. 由于数据已知可分为三类,故采用不同的符号表示不同的类别. 由图 1(a) 可见,SIC 算法由于是随机投影,本文模式随机分散于整个投影的二维平面. 由图 1(b) 可见,MSIC 算法采用主成分分析后的投影方式,各模式投影后已经具有了一定的区别,尤其是第 1 类已经基本上能从另两类中分开. 这种投影方式使得在算法的初始时刻,各模式与其局部环境中的其他模式已经具有了很大的相似性. 实际上相当于 SIC 算法执行了几百甚至是几千个循环后的结果.

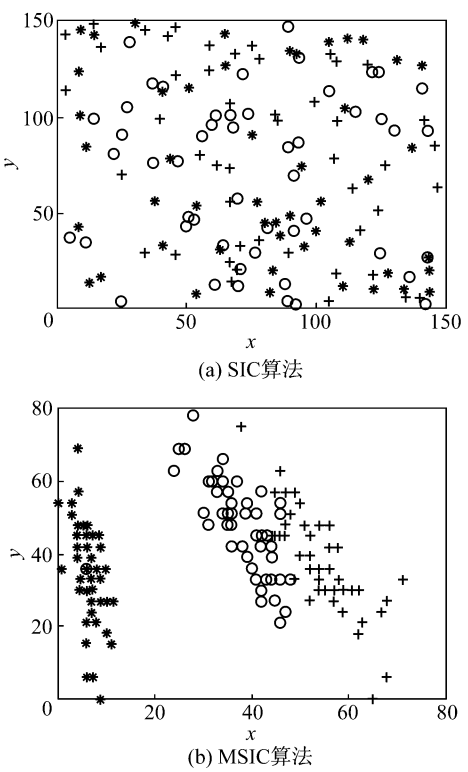

图 1 两种算法初始时刻模式的投影图

Fig. 1 Projections of the patterns in two algorithms

经反复试验,MSIC 一般在经历 8 000～10 000 个循环后,三类就已完全分开. 而 SIC 算法给出的最大循环次数为 60 000. 由于在蚂蚁个数和其他参数基本相同的情况下,算法的运行时间主要取决于循环次数,故 MSIC 算法的运行时间必然会大大缩减. MSIC 算法的聚类结果如图 2 所示. 算法的平均分类正确率为 90.3 %,这与文献[8]中给出 SIC 算法的 90 %相差不大,可见在分类效果基本相同的情况下,MSIC 算法的运行效率有了很大提高.

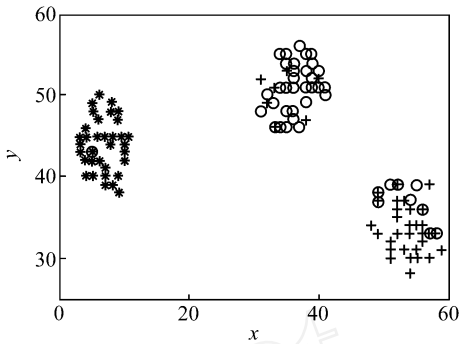

图 2 MSIC 算法的聚类结果

Fig. 2 The clustering results of MSIC algorithm

3 MSIC 算法在设备性能横向比较中的应用

取文献[11]中某大厦中央空调的 43 个 6 维样本数据,去除了文献中提到的 3 个离群点,共有 40 个数据. 6 个属性分别表示空调的送风温度、送风压力、送回风温度差、送风机控制偏差、回风机控制偏差和阀门控制偏差.

MSIC 算法的参数为: cycle_number = 5 000, ant_number = 3, $r = 2$, $F(1) = 4.0$, $k = 0.95$,各模式投影时主成分的处理同上. 经计算,前 2 个主成分保留了原数据的 91.46 %的信息.

由于各模式按主成分的处理值进行投影,故初始投影时,有很多相似的模式就已经聚成一类. 图 3 所示为最终的聚类结果,除了第 16 个点(图中箭头所指)稍微偏离其所属的类别第 8 类外,其他各类的聚类效果非常好. 40 个空调被分为 8 类,每类各 5 个(图中有的模式的坐标发生重叠). 这与文献[10]中的聚类结果完全一致. 这 8 个不同的类分别代表了空调 8 种不同的故障模式.

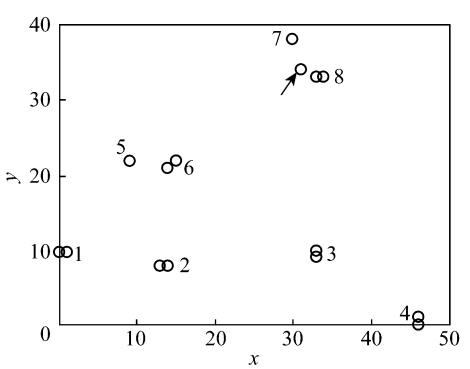

图 3 基于 MSIC 的空调性能聚类分析

Fig. 3 The clustering analysis of air-condition data based on MSIC

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

由此可见,MSIC 算法可以根据数据的特点自组织地得到聚类个数,这和其他需事先给定聚类个数的算法(如 K 均值聚类)相比,具有很大的优越性.

4 结　语

在网络环境下,进行设备性能之间的横向比较是实现智能维护的重要方法.本文提出了一种基于群体智能聚类的设备性能横向比较算法,实现了设备性能的自组织聚类分析.针对 SIC 算法运行效率不高、参数选择缺乏指导等问题,提出了一种改进的 SIC 算法.利用改进后的算法对某中央空调的数据进行了分析,验证了算法的有效性.本文只是对这种基于群体智能的设备性能的聚类分析方法做了仿真实现,今后将在设备的监控主机或网络数据服务器上实现该算法方面做进一步的研究.

参考文献:

[1] Lee J, Kramer B M. Analysis of machine degradation using neural networks based pattern discrimination model [J]. **Journal of Manufacturing Systems**, 1992, 12(3):379 - 387.

[2] Lee J. Measurement of the machine performance degradation using a neural networks model [J]. **International Journal of Computers in Industry**, 1996, 30:193 - 209.

[3] Lin C C, Wang H P. Performance analysis of rotating machinery using enhanced cerebellar model articulation controller neural network [J]. **Computer Industry Engineering**, 1996, 30(2): 227 - 242.

[4] Yan J H, Muammer K. Predictive algorithm for machine degradation detection using logistic regression [A]. **5th International Conference on Managing Innovations in Manufacturing**[C]. Milwaukee, Wisconisn, USA: [s. n.], 2002: 9 - 11.

[5] Dae S H, Lee J. An adaptive Watchdog for monitoring of machine degradation using ART2 (Adaptive Resonance Theory2) [A]. **Proceedings of International Conference on Intelligent Maintenance Systems**[C]. Xi'an, China: The Publishing Company of National University of Defense Technology, 2003: 806 - 810.

[6] Deneubourg J L, Goss S, Frank N, *et al*. The dynamics of collective sorting: Robot-like ants and ant-like robots [A]. **Proceedings of the 1st International Conference on Simulation of Adaptive Behavior: From Animals to Animats** [C]. Cambridge, MA: MIT Press/Bradford Books, 1991:356 - 363.

[7] Lumer E, Faieta B. Diversity and adaptation in populations of clustering ants [A]. **Proceedings of the 3rd International Conference on Simulation of Adaptive Behavior: From Animals to Animats** [C]. Cambridge, MA: MIT Press/Bradford Books, 1994: 501 - 508.

[8] 吴　斌. 群体智能的研究及其在知识发现中的应用[D]. 北京:中国科学院计算技术研究所,2002.

[9] 吴　斌,郑　毅,傅伟鹏. 一种基于群体智能的客户行为分析算法[J]. **计算机学报**,2003,26(8):913 - 918.
WU Bin, ZHENG Yi, FU Wei-peng. A customer behavior analysis algorithm based on swarm intelligence [J]. **Chinese Journal of Computers**, 2003,26(8): 913 - 918.

[10] Fisher R A. Iris plants database[EB/OL]. http://www.ics.uci.edu/～mlearn/ML Repository.html.

[11] 杨欣斌,孙京诰,黄　道. 基于蚁群聚类算法的离群数据挖掘方法[J]. **计算机工程与应用**,2003(9),12 - 13.
YANG Xin-bin, SUN Jing-gao, HUANG Dao. The outlier data mining based on ant swarm clustering alogirthm[J]. **Computer Engineering and Application**, 2003(9),12 - 13.

上海交通大学学报

国家期刊奖百种重点期刊　　百种中国杰出学术期刊

欢迎投稿　　欢迎订阅

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net