改进的全局 K 均值算法及其在啤酒系统中的应用①

张忠厚，赵 龙

(辽宁工程技术大学 理学院, 阜新 123000)

摘 要：K 均值算法存在的问题一直限制其发展，主要问题在于：簇个数的确定、初始聚类中心选择和避免孤立点的问题。针对这些问题进行了改进优化，并把改进后的算法和动态递归模糊神经网络结合一起应用到了啤酒发酵系统当中。神经网络结构复杂，而粒子群算法可以优化全连接网络结构下的各层之间的连接权值和优化网络的拓扑结构。改进的粒子群优化算法也很大程度解决了早熟收敛的问题，有很好的泛化能力，在实际应用中改进的粒子群优化算法原理更简单，参数更少，实现更容易。

关键词：人工智能；粒子群优化算法；K 均值；预测控制；DRFNN

Improved Global K-Means and Its Application in Beer System

ZHANG Zhong-Hou, ZHAO Long

(College of Science, Liaoning Technical University, Fuxin 123000, China)

Abstract：K-means algorithm has been limited by the main questions which are the problems to determine the number of clusters, initial cluster center points of selection and to avoid isolating the problem. To solve these problems the algorithm has been improved in this paper and the paper has applied the improved algorithm and dynamic recurrent fuzzy neural network to the beer fermentation systems. Because of complex neural network structure, the particle swarm optimization algorithm can be used to optimize connected network structure of the connection weights between layers and the network topology. This PSO does not easily trapped local minima and has better generalization ability. At the same time, in practical application the principle of improved PSO algorithm is simple and has less parameter so that it's easier to realize.

Key words：AI; PSO; K-means; predictive control; DRFNN

随数据挖掘在各个领域都应用十分广泛，其中 K 均值算法是比较经典的算法，但是其缺点也限制了自身的发展与应用，近年来众多研究者提出了许多改进的方法和模型，但大部分都是由增加参数，这无疑增加了对算法的人为因素影响，算法性能也受到了影响，也没有解决孤立点的问题，本文应用最小目标函数来解决初始聚类中心的问题，并用改进 PSO 算法搜索最优解，并加入马氏距离来解决孤立点问题。啤酒发酵系统是啤酒生产重要的组成部分，发酵是一个具有高度非线性、时变性和迟滞性的复杂的生物反应过程，同时也是一个放热的过程。各个参数之间又有关联要做到实时控制，对算法的要求很高[1]。

在过去的几十年里，提出了很多种预测控制方法，它们都是建立在预测模型、滚动优化和反馈校正 3 项基本原理基础上的。在实时控制系统中，预测控制部分预测模型的精确度起着非常重要的作用，这决定了预测控制在非线性系统中的应用。对于非线性系统的预测控制问题，可以转化为线性系统预测控制问题，如采用模型线性化方法或者分层递阶优化的方法，也可以直接使用 Hammerstein 模型、Wiener 模型、Volterra 模型、神经网络模型作为预测模型。但 Hammerstein，Wiener，Volterra 模型只能用于特定过程，而神经网络虽然有较好的非线性逼近能力，但仍存在网络拓扑结构难以确定、收敛速度慢、容易陷入局部极小等问题。

① 收稿时间:2011-11-30;收到修改稿时间:2012-01-07

所以在保留神经网络的非线性逼近能力的同时改进网络的拓扑结构是一个很好的突破口。在本文中，提出了改进的 PSO 算法来优化神经网络参数，采用减法聚类简化网络的拓扑结构。从而使神经网络预测更加简洁方便，有更好的精确度[2]。

DRFNN 同时具有递归神经网络和模糊逻辑特点，是一种典型的动态神经网络，它通过与外部的延时反馈可以较好的反映出系统的动态映射关系，因此比前馈神经网络更具有优势，而模糊逻辑具有定性知识表达的能力，所以 DRFNN 在处理参数漂移、强干扰、非线性、不确定性等问题上表现出了较好的性能。动态递归模糊神经网络具有一个特别的递归层，目的是通过前馈神经网络中的附加状态反馈神经元来描述系统的非线性动态行为。递归层的节点数与第 3 层的节点数相同，其用来记忆第 3 层前一时刻的输出值，相当于一步时延算子。可见递归层能存储系统过去的信息，从而增加网络对动态信息的处理能力。

1 系统概述

在标准 PSO 中，粒子的当前位置与个体极值的位置并和全体极值[3]的位置在不断的接近，两个速度更新公式中的速度也在逐渐接近，因此，在更新粒子的速度在迭代后期主要是根据 $\omega \times v$ 变化的[4]。当惯性权重线性下降到小于 1 时，粒子的飞行速度逐渐变小，最后接近于零。因为没有飞行速度现在有粒子出现“惰性”，随着迭代的进行，其他粒子将很快向惰性粒子周围聚集，发生“收敛”的现象，这使得算法过早收敛到局部最优。

如果将粒子群作为一个系统看待那么整个系统内的能量可以保持不变，粒子进入最优点时，速度为零或很小，从能量角度看，必然存在着能量损失。为了使陷入局部最优粒子群自动离开，改进 PSO 是先计算出粒子更新前后的内部能量，算法后期再将损失的能量弥补给发生趋同的粒子。随着能量的恢复，大多数的粒子有足够的能量，那么全局最优粒子组将从粒子群体中分离出来，并在解空间开始重新优化[5]。利用改进后的 PSO 算法优化神经网络参数，选出最优初值，作为训练数据的初始值，使神经网络有更好的预测能力。

预测控制作为一种先进的过程控制方法已被广泛地应用在各种工业生产过程中，它并不要求对模型的结构有先验知识，不必通过复杂的辨识过程便可设计控制系统，在优化过程中不断利用实测信息不断进行反馈校正，在一定程度上克服了模型不确定性的影响，能很好的处理难以建立准确过程模型的过程[6]。文中采用的是 DRFNN 预测控制和 PID 控制结合的串级控制。控制算法如下所示：

Step1 初始化

Step2 设置待优化的 DRFNN 参数作为 MPSO 中粒子的位置向量，代入 DRFNN 预测模型获得 k+1 时刻系统的估计输出 $y(k+d+1)$ 通过偏差修正该估计输出并获得粒子的适应值函数。

Step3 按照 MPSO 算法的规则来寻优，找出粒子的最优适应值。

Step4 更新每个粒子的权重、速度、和位置，检查是否达到了最大的迭代次数或者是评价值小于给定的精度 e；是则退出，当前的最优适应值为优化的参数；否则继续迭代[7]。

Step5 把参数代入 DRFNN，进行模型训练，把这时的参数作为发酵过程中实时测控应用程序中 DRFNN 模型的初始值。

Step6 对非线性系统施加足够的激励信号以得到输入输出数据，构成样本集，使用 DRFNN 对连续样本进行训练并建立 DRFNN 预测模型。

Step7 输出信号控制量和 PID 信号控制量结合控制系统温度值，利用反馈矫正控制量。

采用 DRFNN 预测控制和 PID 控制结合的串级控制对系统进行实时控制，在文献中采用的是测试的方式来寻找合适的模型的参数，这既不能保证找到最优解，也在时间上达不到方便快速的要求[8]。模型建立的好，误差就会减小，这样得出的控制量就是比较优化的可以满足系统在控制过程中的鲁棒性和实时性的要求。这就是本文采用改进的 PSO 优化 DRFNN 参数的目的。结合两者的优势，使控制系统既保证了实时的动态性也简化了模型建立的复杂性。

2 全局K均值算法的改进与优化

在啤酒发酵系统中引入改进后的 K 均值聚类算法，对啤酒发酵系统中的温度数据进行聚类分析，提高了数据预处理的能力，使结果更加清晰，作为预测控制的基础，使整个系统发挥更好的性能。

K 均值聚类算法在很多领域应用十分广泛，因为

它应用简单、快速，但是它也存在着问题，其中比较典型的是初始聚类中心选择的问题直接影响着聚类的效果，所以近年来人们相继对这个问题进行了研究，提出了一些解决的办法，但一般只是针对小型数据库，而且牺牲了一定的速度，本文提出由改进的 PSO 算法进行初始值中心的选取改进，不仅保留了原有算法的速度优势，也解决了初始中心选取的问题。

Zalik 提出的由均方误差函数和不确定信息函数组成的目标函数可以用优化条件的形式表示为

$$F_k = f_k(x) + h_k(x) \tag{1}$$

其中 f_k， h_k 分别为

$$f_k\left(x^1,\cdots,x^k\right)=\frac{1}{m}\sum_{i=1}^{m}\min_{j=1,\cdots,k}\left\|x^j-a^i\right\| \tag{2}$$

$$h_k\left(x^1,\cdots,x^k\right)=\frac{1}{m}E\sum_{i=1}^{k}N_i\left|\log_2\left(p\left(x^j\right)\right)\right|\frac{n!}{r!(n-r)!} \tag{3}$$

以对数来测量信息，最大限度地减少信息的不确定性分配，适当数量的数据集，最小化均方误差函数值，同时使输入数据的聚类成为可能，这两个函数的总和的全局最小对应为实际群体。a^i 表示数据样本集 A 的第 i 个样本， f_k 表示的是均方误差函数， h_k 表示是不确定信息函数。E 为一个常数，用来选择测量单位。E 的值应该在坐标范围之内，但主要的影响因素在于点的距离， E 的值需要从实验中得到。 r 是小于样本个数大于 0 的随机数。 N 为数据样本个数，$p(x^i)$表示输入数据落入聚类中心为 x^i 的子集中的概率。利用最小目标函数作为条件，然后加入 PSO 算法搜索出最优结果作为初始聚类中心。由于 f_k 是随着 k 值变化的增大而减小，序列 h_k 随 k 值的增大而增大。因此可以通过最小化目标函数来确定聚类中心个数。为此引进方程

$$d_{k-1}^i=\min\{\left\|x^1-a^i\right\|^2,\cdots,\left\|x^{k-1}-a^i\right\|^2\} \tag{4}$$

$$b_j=\sum_{i=1}^{m}\max\left\{0,d_{k-1}^i-\left\|a^j-a^i\right\|\right\} \tag{5}$$

一个使 b_j 达到最大值的点 a^j 是第 k 个初始聚类中心的最佳选择，然后引进马氏距离，马氏距离不因样本的特征向量值的测量单位、数量级上的差异而导致距离计算上的差异，所以采用马氏距离剔除数据中的错误数据，作为避免孤立点的问题。马氏距离作为 K 均值算法中样本与聚类中心之间的最小距离准则当中的距离，使其对样本参考量的影响降到最低。使用 PSO 算法搜索出最小化目标函数的最优解，定义各个粒子的值为初始样本之间的马氏距离，搜索孤立点，然后剔除孤立点粒子，重新搜索。

采用 Krista 所提出的目标函数，这个目标函数的特征是样本集区强烈的吸引聚类中心，为了给由输入数据点所组成的簇最大的信息量，每个聚类中心驱使所有其他的聚类中心远离输入数据点。这样就能够移出多余的聚类中心，使其远离数据集，而相应的簇可以忽略不计。

3　仿真

基于对啤酒发酵过程，它需要保持恒定的压力的要求，温度应实时调整根据不同的控制要求。设置系统步骤为：采集数据，对数据进行预处理和清洗，采用改进的全局 K 均值算法对数据进行聚类分析，使数据规格化，再利用优化后的神经网络对数据进行分析进而对系统做出预测，针对预测数据做出调节。设置控制时域为 2 和预测时域为 9，设置一个采样周期为 10s[9]，仿真误差结果为图 1 所示。从图中，改进 PSO 算法 DRFNN 具有更好的可预见性，使温度控制具有更好的动态，鲁棒性和稳定性。这个控制系统不仅具有响应速度快和超调量小，也能满足对给定的温度和限制系统超调量小，运行平稳的特点的波动的要求。在常规 PID 调节模式，系统的控制性能并不令人满意。PID 是一般线性模型的基础上，并不能起到很好的作用在大时滞控制系统。结果表明，改进的 PSO 算法在 DRFNN 参数优化中的应用是可行的。它是可行和实际的结合预测控制与 PID 控制啤酒发酵温度。

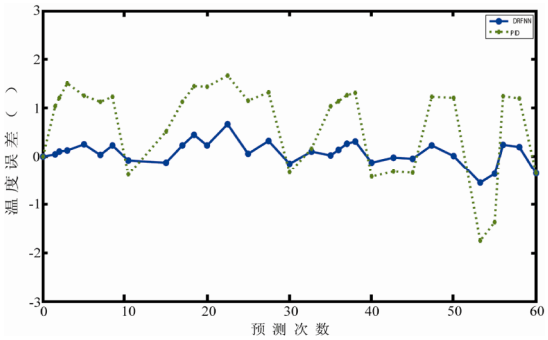

图 1　仿真结果

(下转第 239 页)

本地库是否要适应多核和众核的环境；Java 类库已经在 Web Service 编程方面作出了积极的努力，推动了服务计算技术的发展；Java 字节码文件是否增加新的内容以适应新技术的发展呢；Java 在对象数据库开发方面如何提供支持；由于 Java 技术框架的特点，在高性能计算方面，Java 一直没有好的表现。然而 Java 在 HPC 方面的努力一直没有停止过。

7 小结

Java 技术是目前流行的、最广泛采用的软件开发技术，但很多人只是从 Java 程序设计语言的角度理解 Java 技术，缺乏对 Java 技术框架的全面认识。本文针对 Java 的整体技术框架进行了概述，阐述了 Java 程序设计语言、Java 虚拟机、Java 库、Java 字节码文件之间的关系，最后在新兴技术不断涌现的今天，对 Java 技术的发展趋势进行了探讨。

参考文献

1 Java 虚拟机规范.http://java.sun.com/docs/books/jvms/second_edition/html/VMSpecTOC.doc.html

2 Gosling J, Joy B, Steele G, Bracha G. Java™ Language Specification. The 3rd Edition. Prentice Hall, Jun 14, 2005.

3 Horstmann CS, Cornell G.叶乃文,等译.JAVA2 核心技术卷 I:基础知识(原书第 7 版).北京:机械工业出版社,2006.

4 Horstmann CS, Cornell G.陈昊鹏,等译.JAVA2 核心技术卷 II:高级特性(原书第 7 版).北京:机械工业出版社,2006.

(上接第 196 页)

4 结语

在本文中，改进的全局 K 均值算法解决了聚类分析时的问题，不仅可以利用最小目标函数来解决初始中心问题，还可以避免孤立点的问题。马氏距离有效地消除错误数据的样本数据，从而减少了计算量。因为动态模糊神经网络是不容易获得的梯度信息，网络培训时间太长，结构复杂等缺点，减法聚类用于简化网络结构，并把改进 PSO 算法用于优化网络参数，避免经验的盲目性和随意性。仿真结果表明，它优于传统的动态递归模糊神经网络，它需要更少的隐藏节点和参数，具有较高的预测精度和实用性。

参考文献

1 薛尧予,王建林,于涛,赵利强.基于改进 PSO 算法的过程模型参数估计.仪器仪表学报,2010,31(1):178−182.

2 Shivakumar RL. Implementation of an Innovative Bio Inspired GA and PSO Algorithm for Controller design considering Steam GT Dynamics. Journal of Computer Science Issues, January 2010,24(2):132−140.

3 武俊峰,艾岭.一种基于改进算法的模糊模型辨识.哈尔滨理工大学学报,2010,20(5):165−170.

4 穆朝絮,张瑞民,孙长银.基于粒子群优化的非线性系统最小二乘支持向量机预测控制方法.控制理论与应用,2010,27(2):164−168.

5 王博,孙玉坤.基于数据场聚类的模糊神经网络在发酵过程中的应用.仪器仪表学报,2009,30(5):944−948.

6 Xue YY, Wang JL, Yu T, Zhao LQ. Parameter estimation of fermentation process model based on an improved PSO algorithm. Chinese Journal of Scientific Instrument, 2010, 31(1):178−182

7 田雨波.混合神经网络技术.2nd ed.北京:科学出版社,2009.148−160.

8 吕奕清,林锦贤.基于 MPI 的并行 PSO 混合 K 均值聚类算法.计算机应用,2011,32(1):112−118.

9 李国勇.智能控制及其 MTLAB 的实现.3rd ed.北京:电子工业出版社,2007.156−188.