第41卷 第3期 太 阳 能 学 报 Vol. 41, No. 3
2020年3月 ACTA ENERGIAE SOLARIS SINICA Mar., 2020

文章编号:0254-0096(2020)03-0008-07

# 基于并行模糊*C*-均值聚类的风电机组发电机故障诊断研究

孙鹤旭[1]，孙泽贤[1]，张靖轩[2]

（1. 河北工业大学控制科学与工程学院，天津 300130；2. 华北理工大学电气工程学院 唐山 063000）

**摘 要**：提出一种基于并行化的改进模糊*C*-均值聚类的风电机组发电机故障诊断方法。首先通过邻近聚类算法确定数据集可能的最大类簇数$C_{max}$，以$C_{max}$为上限，利用改进的模糊*C*-均值算法，以BWI(between-within index)指标为聚类有效性判别指标，提出新的确定最佳聚类数的方法。并结合Spark内存处理技术，将其应用在风电机组发电机的故障诊断中，通过UCI机器学习数据库数据集以及风电监测实时数据的实验测试，表明该算法不仅能准确判别发电机的故障模式，并且能更好地处理电力系统的海量数据。

**关键词**：故障诊断；聚类；风电机组；发电机；风电监测；Spark

**中图分类号**：TM914 **文献标志码**：A

## 0 引 言

随着近年来风电机组的单机容量不断增大，风电场设备故障频频发生，严重影响了机组的运行效益及电网安全运行[1]。但针对不同厂商、不同型号的风电机组，在保证以秒级周期采集数据的前提下还需面对数据结构、数据接口、存储方式等数据库体系的差异性，因此构成了多源、异构、量大的风电监测大数据。因此对风电机组设备实现有效的故障诊断就显得尤其重要。

目前，已有学者对风电机组的故障诊断进行了深入研究。文献[2]利用粒子群算法优化神经网络，提取齿轮箱的故障特征。文献[3]借助粗糙集的动态层次聚类提取出故障诊断规则。文献[4]采用模糊*C*-均值算法实现了对光伏阵列故障模式的有效识别。但上述方法都是在单机环境小数量级下进行的，面对异构的海量数据，算法的性能并不能得到有效保证。文献[5]设计出基于Hadoop的MP Apriori算法，在电网数据中挖掘连锁故障各站点之间的关联。文献[6]结合Spark和Storm处理技术搭建了故障诊断与预警模型。文献[7]提出一种基于Spark的并行*C*-均值算法辨识不良数据的新方法。文献[8]在Spark平台下实现了模糊*C*-均值算法的并行计算。文献[9]采用并行化后的粒子群算法优化支持向量机参数，进而实现短期负荷预测。文献[10]利用Spark平台分割全部数据并搭建回归模型提出新的电力负荷预测算法。

本文通过分析风电机组故障类型和故障特征之间的内部关联，给出基于Spark框架的风电机组故障诊断模型(以发电机故障诊断为例)。本文结合邻近聚类算法和最大最小距离算法的思想，确定了聚类数的搜索范围，并提出一种新的聚类有效性判别指标——BWI指标，进而获取数据的有效分类；最后，结合Spark内存处理，将算法应用在发电机的故障诊断中，实验证明了算法的有效性。

## 1 改进的模糊*C*-均值聚类算法

### 1.1 模糊*C*-均值聚类算法

传统的模糊*C*-均值(fuzzy *C*-means algorithm，FCM)算法是一种基于划分的聚类算法[11]，它的思想是异类的类间距离最大及同类的类内距离最小，从而定量的判断研究对象之间的隶属关系，以达到对样本数据按照类别自动分类的目的。

设样本集为$\boldsymbol{X}=\{x_1,x_2,\cdots,x_N\}$，聚类中心为$\boldsymbol{V}=\{v_1,v_2,\cdots,v_C\}$，第$j$个样本对第$i$个聚类中心的隶属度为$u_{ij}$，FCM的聚类优化模型为：

**收稿日期**：2017-05-08

**基金项目**：河北省科技计划(17214304D)；天津市科技支撑项目(14ZCDZGX00818)

**通信作者**：孙泽贤(1991—)，男，博士研究生，主要从事新能源发电技术方面的研究。shx13682168380@sina.com

$$\begin{cases} J(\boldsymbol{U},\boldsymbol{V})=\sum_{i=1}^{C}\sum_{j=1}^{N}u_{ij}^{m}d_{ij}^{2} \\ \text{s.t.}\sum_{i=1}^{C}u_{ij}=1,\quad 0\leqslant u_{ij}\leqslant 1 \end{cases} \tag{1}$$

式中，$\boldsymbol{J}$ ——目标优化函数；$\boldsymbol{U}$ ——模糊隶属度集合；$\boldsymbol{V}$ ——聚类中心集合；$m$ ——决定聚类结果的平滑系数；$d_{ij}$ ——第 $j$ 个样本与第 $i$ 个聚类中心的欧氏距离。

运用拉格朗日乘子算法，求导后得到必要条件为：

$$V_i=\frac{\sum_{j=1}^{N}u_{ij}^{m}x_j}{\sum_{j=1}^{N}u_{ij}^{m}} \tag{2}$$

$$u_{ij}=\frac{1}{\sum_{k=1}^{C}\left(\frac{d_{ij}}{d_{kj}}\right)^{\frac{2}{m-1}}} \tag{3}$$

式中，$V_i$ ——第 $i$ 个聚类中心；$u_{ij}$ ——第 $j$ 个样本隶属于第 $i$ 个类别的隶属度；$k$ ——聚类中心的索引号。

FCM 算法根据上述 2 个必要条件，结果输出 $C$ 个聚类中心以及 1 个（$c\cdot n$）的隶属度矩阵 $\boldsymbol{U}$ 。通过比较矩阵中最大隶属度，从而判断样本的归类情况。

## 1.2　基于Spark的并行FCM算法

### 1.2.1　Spark 架构和分布式数据集 RDD

Spark 是目前最为流行的分布式计算系统，通过引入弹性分布式数据集（resilient distributed dataset，RDD）构建多元化的大数据处理体系。Spark 是一种基于内存计算、缓存的计算系统，减少了大量的磁盘 IO 操作，极大提高了需要多次迭代的复杂计算的处理速度。

集群资源管理服务器（cluster manager）把任务控制节（driver program）划分好的任务集分发到各个节点（worker node），每个节点上的进程（executor）执行相应的任务（task），Spark 的运行框架如图 1 所示。

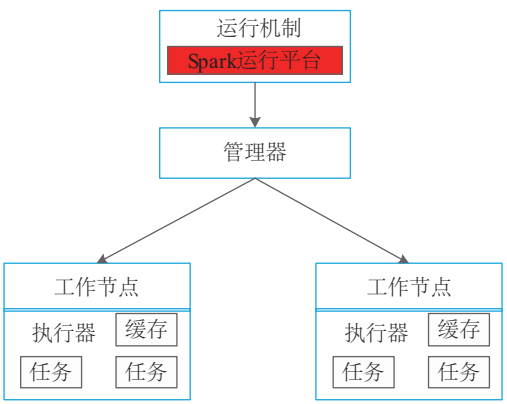


图 1　Spark 运行结构

Fig. 1　Architecture of Spark

### 1.2.2　新的聚类有效性指标——BWI 指标

聚类结果的有效性分析就是评价聚类结果优劣的过程。聚类的目标是呈现“类内相似、类间相异”的特性。目前已提出一些有效性指标，由于这些指标自身的缺陷，一般难以找到正确的最佳聚类数。鉴于这种情况，本文设计一种新的聚类有效性指标——BWI 指标。

定义 1：$K=(X,\ A)$ 表示聚类空间，$X=(x_1,x_2,x_3\cdots x_n)$ 为样本对象，假设样本对象 $X$ 被分为 $C$ 类，定义第 $i$ 个样本的类内距离为 $w(j,i)$ 该样本到同类中其他样本距离的平均值，即：

$$w(j,i)=\frac{1}{n_j-1}\sum_{q=1}^{n_j}\left\|x_q^{(j)}-x_i^{(j)}\right\|^2 \tag{4}$$

式中，$x_q^{(j)}$ ——所属第 $j$ 类中的第 $q$ 个样本；$n_j$ ——第 $j$ 类的样本数量。

定义 2：$K=(X,\ A)$ 表示聚类空间，$X=(x_1,x_2,x_3\cdots x_n)$ 为样本对象，假设样本对象 $X$ 被分为 $C$ 类，定义样本的类间距离 $b(j,i)$ 为不同类间聚类中心距离的最小值与类间聚类中心距离的平均差异度之和，即：

$$b(j,i)=\min_{i\neq j}\left(\left\|v_i-v_j\right\|^2\right)+var\left(\left\|v_i-v_j\right\|^2\right) \tag{5}$$

式中，$v_i$ ——第 $i$ 个簇质心；$v_j$ ——第 $j$ 个簇质心。

平均差异度 $var\ (\cdot)$ 为簇质心距离值和与聚类中心数量的比值，如式（6）所示：

$$var\left(\left\|v_i-v_j\right\|^2\right)=\frac{\sum_{i=1}^{c-1}\sum_{j=i+1}^{c}\left\|v_i-v_j\right\|^2}{c} \tag{6}$$

定义 3：$K=(X,\ A)$ 表示聚类空间，$X=(x_1,x_2,x_3\cdots x_n)$ 为样本对象，假设样本对象 $X$ 被分为 $C$ 类，定义第 $j$ 类的第 $i$ 个样本的类间类内指标（between-within index，BWI）指标为类内距离和类间距离的比值，即：

$$\begin{aligned} BWI(j,i)&=\frac{b(j,i)}{w(j,i)} \\ &=\frac{\left(\frac{1}{n_j-1}\sum_{q=1}^{n_j}\left\|x_q-x_j\right\|^2\right)}{\min_{i\neq j}\left(\left\|v_i-v_j\right\|^2\right)+var\left(\left\|v_i-v_j\right\|^2\right)} \end{aligned} \tag{7}$$

BWI 指标基于数据样本的几何结构，对聚类结果进行有效性分析。BWI 指标值越大，说明对应样本的聚类效果越好，整体样本的 BWI 指标的平均值反映了数据集的聚类效果，平均值越大，聚类效果越好。BWI 指标的最大平均值所对应的 $C$ 即为最佳聚类数，由此得到式（8）、式（9）：

$$\text{avg}BWI(C)=\frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}BWI(j,i) \tag{8}$$

$$Copt=\arg\max\left\{\arg\underset{2\leqslant k<n}{BWI(C)}\right\} \tag{9}$$

1.2.3 基于近邻传播聚类算法确定 $C_{max}$

$K=(X, A)$ 表示聚类空间，$X$ 为样本集，$A$ 为属性集，引入近邻传播聚类算法，通过计算样本的可信度和可用度确定样本的类代表，最后通过初步的聚类结果确定 $C_{max}$。

基于近邻传播聚类算法确定 $C_{max}$ 描述如下：

1)初始化，计算样本间的相似度矩阵 $[s(i,k)]_{N\cdot N}$，其中 $N$ 为样本数量；设定相似度矩阵对角线元素 $s(k,k)$ 为吸引度中值 $pm$；设定初始可信度 $r(i,k)$ 和初始可用度 $a(i,k)$ 为 0；设定振荡因子 $lam$ 初值。

2)AP 算法迭代过程为：① 计算新的可信度和可用度。② 根据 $\arg\max_k\{r(i,k)+a(i,k)\}$ 原则确定样本 $x_i$ 的类中心样本。③ 判断算法是否满足迭代停止条件：迭代次数超过最大迭代次数，及类中心连续几步迭代过程保持稳定。

3)迭代结束，得到聚类结果。

4)更改不同的振荡因子，可能会得到不同的聚类结果，选取簇数量的最大值作为 $C_{max}$。

1.2.4 改进的 FCM 算法思想

FCM 算法通常是随机选取 $C$ 个样本作为初始聚类中心，对于结构复杂的数据集，聚类结果会随着初始聚类中心的变化而波动。因此，本文结合聚类中心的初始化和 BWI 有效性指标来确定最优聚类结果。改进的 FCM 算法步骤为：

1)确定聚类数的范围 $[C_{min}, C_{max}]$。其中，$C_{min}$ 取值为 2，$C_{max}$ 由 1.2.3 节方法得出。

2)选取所有初始聚类中心。基于最大最小距离算法选取 $C_{min}$ 个样本作为初始聚类中心以后每增加一个聚类数，在保持上一组初始聚类中心不变的情况下，按照最大最小距离算法的计算出新的初始聚类中心，另外，按照此方法，聚类数是提前已知的，因此避免结果受到比例系数的影响。

3) for $C=C_{min}$ to $C_{min}$

① 按照第 2)步方法初始化 $C$ 个初始聚类中心；

② 运行模糊 $C$ 均值聚类算法，更新隶属度矩阵 $U^k$ 和聚类中心 $C^k$；

③ 检查是否满足算法终止条件，如不满足，返回②；

④ 根据聚类结果计算 BWI 指标，返回 3)；

4)比较 BWI 指标值，选取指标值最大的聚类结果作为最优聚类结果。

1.2.5 基于 Spark 的改进 FCM 算法的并行化实现

利用 Spark 实现 FCM 算法的并行化计算，主要采用“map”、“reduce”算子，在迭代的过程中，先用“map”计算样本对应的隶属度矩阵，再用“reduce”构建新的聚类中心。不同于 Hadoop 的 MapReduce 计算框架，Spark 中的迭代计算都是在内存中对 RDD 完成，无需与磁盘交互。

基于 Spark 的 FCM 算法并行化实现分两部分：1)首先读取已预处理过的文件，创建新的 RDD，并执行 cache 操作对 RDD 数据进行缓存。在 Map 阶段计算样本数据的均值，然后在 Reduce 过程中根据最大最小距离算法得到 $k$ 个初始聚类中心；2)通过 Map 操作及上一步生成的初始聚类中心得到隶属度矩阵，Reduce 操作根据隶属度矩阵，计算新的聚类中心。与 MapReduce 相比，Spark 将需要迭代计算的数据集定义为 RDD，并将其分布在不同的节点上，由节点所在的 task 完成迭代计算，节省了大量的磁盘读取时间。

## 2 基于并行FCM聚类的发电机故障诊断

### 2.1 选取故障参数

为准确表征发电机的故障特性，选取的特征参数必须能高效地刻画发电机运行状态。在选取特征参数时要结合专家知识，同时考虑风力机监测的采集点分布。风力机监测系统保存发电机相关参数，当发电机发生故障时某些遥测参数就会升高或降低。本文选取表 1 所示参数 $a$~$h$ 作为特征参数，为获得故障样本数据集，按上述选取的特征参数读取发电机实时监测数据，表 1 为部分发电机实时数据。

**表1 发电机实时数据(部分)**

Table 1 Real-time data of alternator

| 特征参数/℃ | 采集时间点/min | | | | |
|---|---|---|---|---|---|
| | 0 | 15 | 30 | 45 | 60 |
| $a$ | 39 | 41 | 45 | 50 | 50 |
| $b$ | 38 | 41 | 45 | 49 | 48 |
| $c$ | 22 | 23 | 25 | 26 | 25 |
| $d$ | 29 | 30 | 31 | 32 | 32 |
| $e$ | 34 | 35 | 39 | 41 | 40 |
| $f$ | 0 | 0 | 0 | 0 | 0 |
| $g$ | 11 | 12 | 13 | 11 | 10 |
| $h$ | 22 | 23 | 23 | 23 | 23 |

表 1 中，$a$ 为发电机绕组温度 1；$b$ 为发电机绕组温度 2；$c$ 为发电机前轴承温度；$d$ 为发电机后轴承温度；$e$ 为发电机冷却空气温度；$f$ 为环境温度；$g$ 为机舱温度；$h$ 为顶控柜温度。

### 2.2　基于并行FCM聚类的发电机故障诊断

基于并行 FCM 聚类的发电机故障诊断过程主要分为以下 2 步：

1)运用并行的 FCM 聚类算法将故障样本聚集成最优的 $C$ 种故障模式。并行的 FCM 聚类方法用于发电机的故障诊断是将已有的故障记录分成最优的 $C$ 类，也就是建立最优的聚类中心。发电机的 FCM 聚类故障诊断本质上就是依据最大隶属度原则对实时采集的状态数据进行识别，进而得到发电机最优聚类中心及其故障模式。

2)为确定待诊断样本的所属模式，需要判断测试样本与上一步计算得出的各个聚类中心的隶属度大小。这里采用式(3)作为隶属度函数进行故障的模式识别。图 2 为发电机的并行 FCM 故障诊断流程。

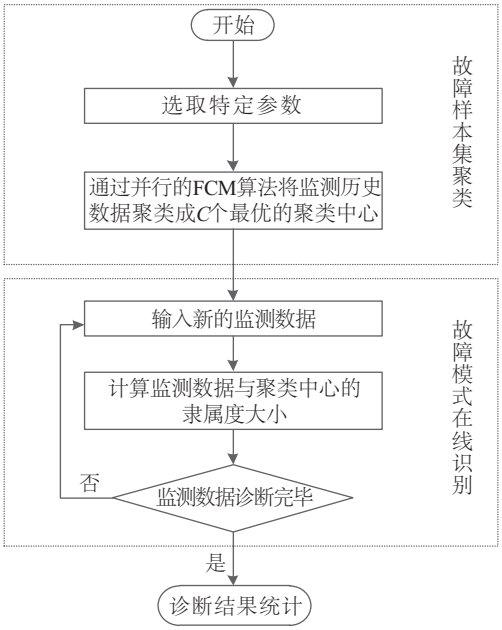


图2　风电机组的并行FCM故障诊断

Fig. 2　Fault diagnosis for wind turbine based on parallel FCM

## 3　实验结果与分析

本文算法使用 Spark 编程环境实现，为模拟风电机组监测大数据下故障诊断情况，搭建风电机组故障诊断实验平台，实验环境由 4 台联想服务器构建故障大数据处理平台。

各节点的硬件配置：CPU 型号 E5-2630 10-core，内存 16 GB，网络带宽 1000 Mbit/s。其中一台服务器作为主控制点 JobTracker，其余 3 台服务器作为 TaskTracker 节点，Hadoop 采用 2.6.2 版本，Spark 的版本是 1.6.0 版本，Java 版本是 1.8.0。为测试算法的有效性，分别在 UCI 标准数据集和风场现场监测数据上进行实验。文中所用的 UCI 数据集描述见表 2。

**表2　UCI数据集描述**

Table 2　Description of datasets from UCI

| 数据集 | 样本数 | 属性个数 | 类别数 |
|---|---|---|---|
| Iris | 150 | 4 | 3 |
| Haberman | 306 | 4 | 2 |
| Bupa | 345 | 7 | 2 |
| Pima-indians-diabetes | 768 | 9 | 2 |

### 3.1　确定$C_{max}$的实验仿真

将 1.2.3 节基于 AP 算法确定 $C_{max}$ 的方法在表 2 的 4 个 UCI 数据集上运行，表 3 为其确定的最大 $C_{max}$ 和经验值的对比。

**表3　各数据集对应的$C_{max}$**

Table 3　Corresponding $C_{max}$ on datasets

| 数据集 | $C_{max}$ | 经验值 $\mathrm{int}\sqrt{n}$ |
|---|---|---|
| Iris | 4 | 12 |
| Haberman | 3 | 17 |
| Bupa | 6 | 18 |
| Pima-indians-diabetes | 8 | 27 |

从表 3 的结果中可看出基于 AP 算法确定的最大簇类数比通常采用的经验值 $\mathrm{int}\sqrt{n}$ 小很多，因此，本文采用 AP 算法确定的 $C_{max}$ 作为上限，大大减小了搜索范围。由表 2、表 3 可知，对于确定大数据集的 $C_{max}$ 依然有很好的效果。

### 3.2　确定最佳聚类数的实验仿真与分析

对表 2 的 4 个 UCI 数据集以表 3 确定的 $C_{max}$ 作为搜索上限，分别运行 1.3.4 节所提出的算法确定最佳聚类数算法。表 4 给出了不同 $C$ 值所对应的不同平均 BWI 指标值，带有下划线的指标值所对应的聚类数为算法确定的最佳聚类数。从表 4 的实验结果来看，本文算法对表 4 中数据集得到的最佳聚类数与实际最佳聚类数完全相同，证明了本文算法的实用性。

**表4 UCI数据集上确定最佳聚类数的BWI指标值**

Table 4 Score of BWI of our althorithm to determine optimal number of clusters on UCI datasets

| $K$值 | BWI指标值 | | | |
|---|---|---|---|---|
| | Iris | Haberman | Bupa | Pima |
| 1 | 0 | 0 | 0 | 0 |
| 2 | 0.4289 | 0.6174 | 0.3467 | 0.4192 |
| 3 | 0.5111 | 0.5794 | 0.3234 | 0.3035 |
| 4 | 0.4694 | 0.4052 | 0.2849 | 0.2572 |
| 5 | — | — | 0.2569 | 0.2224 |
| 6 | — | — | 0.2519 | 0.2338 |
| 7 | — | — | — | 0.2130 |
| 8 | — | — | — | 0.1982 |

为进一步证明本文提出算法的有效性，对表2所示4个UCI数据集以表3中基于AP算法所确定的$C_{max}$为上限，运行改进后的模糊$C$均值算法，并使用常用的聚类有效性指标，即BWP、DB确定最佳聚类数。BWP、DB确定的最佳聚类数结果如表5所示，带有下划线的指标值所对应的聚类数为算法确定的最佳聚类数。

从表4、表5实验结果可知，聚类有效性指标BWI在4种数据集上均能得到正确的簇类数，并且BWP、DB指标除了iris数据集自身数据交错现象而确定错误的簇类数，其余数据集的结果均与实际情况相同，这也进一步证实了本文提出的改进FCM算法的有效性以及BWI指标的准确性。

**表5 UCI数据集上确定最佳聚类数的BWP & DB指标值**

Table 5 Score of BWP & DB of our althorithm to determine optimal number of clusters on UCI datasets

| $K$值 | BWP指标值 | | | | DB指标值 | | | |
|---|---|---|---|---|---|---|---|---|
| | Iris | Haberman | Bupa | Pima | Iris | Haberman | Bupa | Pima |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | −0.534 | −0.261 | −0.442 | −0.409 | 1.205 | 1.561 | 1.545 | 1.440 |
| 3 | −0.647 | −0.399 | −0.572 | −0.671 | 1.379 | 1.775 | 1.944 | 1.639 |
| 4 | −0.647 | −0.494 | −0.684 | −0.741 | 1.546 | 1.601 | 1.948 | 2.038 |
| 5 | — | — | −0.706 | −0.775 | — | — | 1.907 | 2.483 |
| 6 | — | — | −0.706 | −0.776 | — | — | 1.765 | 2.027 |
| 7 | — | — | — | −0.776 | — | — | — | 1.685 |
| 8 | — | — | — | −0.776 | — | — | — | 1.450 |

### 3.3 发电机的故障诊断实验与分析

实验数据集为某风电场的前置采集服务器和实时库服务器获取的实际运行数据，统一整合为.txt格式。本文从故障样本中随机选择一定数量的发电机故障记录和正常运行数据测试算法的故障诊断率，实验结果如表6所示。

**表6 发电机故障诊断结果**

Table 6 Diagnosis results of alternator fault

| 实验环境 | 准确率/% | 误报率/% | 运行时间/s |
|---|---|---|---|
| 单机环境 | 85.2 | 14.8 | 522 |
| Hadoop环境 | 88.4 | 11.6 | 175 |
| Spark环境 | 90.6 | 9.4 | 84 |

实验分别基于单机环境、Hadoop环境以及Spark环境下测试。单机环境所取得的故障诊断准确率和Hadoop、Spark的执行结果基本一致。并且在数据及较大的前提下，通过对比运行时间，Spark并行性能要优于Hadoop平台。

## 4 结 论

针对风电大数据环境下风电机组状态监测中数据实时处理的需求，本文设计实现了基于Spark的风电机组故障诊断模型。首先通过结合AP算法与最大最小距离算法确定初始聚类中心，减小了结果的波动性；提出BWI聚类有效性指标，提高了聚类的准确率；最后，利用Spark内存处理技术实现FCM算法的并行化，通过训练得到故障诊断的模型，加速了风电机组故障诊断的处理效率。实验结果表明，本文提出的方法能够较好地保证诊断率，并同时能够满足电力系统的实时性要求。

## [参考文献]


[1] 宋磊. 双馈异步风电机组状态监测与故障诊断系统的研究[D]. 保定：华北电力大学，2015.
SONG L. Research on conidtion monitoring and fault diagnosis system for double-fed asynchronous wind turbine[D]. Baoding: North China Electric Power University, 2015.

[2] 龙泉，刘永前，杨勇平. 基于粒子群优化BP神经网络的风电机组齿轮箱故障诊断方法[J]. 太阳能学报，2012，33(1)：120-125.
LONG Q, LIU Y Q, YANG Y P. Fault diagnosis method of wind turbine gearbox based on BP neural network trained by practice swarm optimazation algorithm [J]. Acta energiae solaris sinica, 2012, 33(1): 120-125.

[3] 李宁，王李管，贾明滔，等. 基于信息融合理论的风机故障诊断[J]. 中南大学学报(自然科学版)，2013，44(7)：2861-2866.
LI N, WANG L G, JIA M T, et al. Faults intelligent diagnosis system for fan based on information fusion[J]. Journal of Central South University (science and technology), 2013, 44(7): 2861-2866.

[4] 毕锐，丁明，徐志成，等. 基于模糊C均值聚类的光伏阵列故障诊断方法[J]. 太阳能学报，2016，37(3)：730-736.
BI R, DING M, XU Z C, et al. Pvarray fault diagnosis based on FCM[J]. Acta energiae solaris sinica, 2016, 37(3): 730-736.

[5] 曲朝阳，朱莉，张士林. 基于Hadoop的广域测量系统数据处理[J]. 电力系统自动化，2013，37(4)：92-97.
QU C Y, ZHU L, ZHANG S L. Data processing of Hadoop-based wide area measurement system[J]. Automation of electric power system, 2013, 37(4): 92-97.

[6] 张少敏，毛冬，王保义. 大数据处理技术在风电机组齿轮箱故障诊断与预警中的应用[J]. 电力系统自动化，2016，40(14)：129-134.
ZHANG S M, MAO D, WANG B Y. Application of big data processing technology in fault diagnosis and early warning of wind turbine gearbox[J]. Automation of electric power system, 2016, 40(14): 129-134.

[7] 孟建良，刘德超. 一种基于Spark和聚类分析的辨识电力系统不良数据新方法[J]. 电力系统保护与控制，2016，44(3)：85-91.
MENG J L, LIU D C. A new method for identifying bad data of power system based on Spark and clustering analysis[J]. Power system protection and control, 2016, 44(3): 85-91.

[8] BHARILL N, TIWARI A, MALVIYA A. Fuzzy based scalable clustering algorithms for handling big data using apache spark[J]. IEEE transactions on big data, 2016, 2(4): 339-352.

[9] 王保义，王冬阳，张少敏. 基于Spark和IPPSO_LSSVM的短期分布式电力负荷预测算法[J]. 电力自动化设备，2016，36(1)：117-122.
WANG B Y, WANG D Y, ZHANG S M. Distributed short-term load forecasting algorithm based on Spark and IPPSO_LSSVM[J]. Electric power automation equipment, 2016, 36(1): 117-122.

[10] 马天男，牛东晓，黄雅莉，等. 基于Spark平台和多变量L2-Boosting回归模型的分布式能源系统短期负荷预测[J]. 电网技术，2016，40(6)：1642-1649.
MA T N, NIU D X, HUANG Y L, et al. Short-term load forecasting for distributed energy system based on Spark platform and mutil-variable L2-Boosting regression model[J]. Power system technology, 2016, 40(6): 1642-1649.

[11] 曹愈远，张建，李艳军，等. 基于模糊粗糙集和SVM的航空发动机故障诊断[J]. 振动、测试与诊断，2017，37(1)：169-173.
CAO Y Y, ZHANG J, LI Y J . Aero-engine fault diagnosis based on fuzzy rough set and SVM[J]. Journal of vibration, measurement & diagnosis, 2017, 37(1): 169-173.

# FAULT DIAGNOSIS OF WIND TURBINE ALTERNATOR BASED ON PARALLEL FUZZY *C*-MEANS CLUSTERING ALGORITHM

Sun Hexu[1], Sun Zexian[1], Zhang Jingxuan[2]

(1. *Institute of Control Science and Engineering*, *Hebei University of Technology*, *Tianjin* 300130, *China*;
2. *College of Electrical Engineering*, *North China University of Science and Technology*, *Tangshan* 063000, *China*)

**Abstract**: A new method based on improved FCM(fuzzy *C*-means) on fault diagnosis for wind turbine is proposed. First, the affinity propagation clustering algorithm is used to determine the max number of clusters. Then the improved FCM and the criterion of BWI(Between-Within Index)are combined to determine the optimal number of clusters of a dataset. The parallel FCM based on Spark is used to detect the fault in the wind turbine. The improved FCM algorithm are evaluated by studying the UCI datasets and camparing with the real-time data from the experiments, the results show that the method can effectively improve the accuracy fault diagnosis and can better process massive data in power system.

**Keywords**: fault diagnosis; clustering; wind turbine; electrical generator; wind power monitoring; Spark