DOI: 10.7667/PSPC180399

基于并行变量预测模型的变压器故障诊断及优化研究

马利洁，朱永利，郑艳艳

(新能源电力系统国家重点实验室(华北电力大学)，河北 保定 071003)

摘要：针对传统变压器故障诊断方法存在小样本问题下分类效果差、海量监测数据的识别效率低下等问题，提出基于 Spark 计算框架的并行化变量预测模型。首先采用 HDFS 作为内存式存储系统，面向行存储的 RowMatrix 作为分布式矩阵存储结构，利用广播变量、调整分区数进行并行度优化。其次训练 4 种数学模型获取故障类型的最佳模型及相关参数完成故障诊断。实验结果表明，并行变量预测模型识别精度高于支持向量机，计算效率优于单机环境，对高维特征向量有较好的适应性。

关键词：故障诊断；小样本；变量预测模型；Spark 计算框架；内存式存储

Research on transformer fault diagnosis and optimization based on parallel variable prediction model

MA Lijie, ZHU Yongli, ZHENG Yanyan

(State Key Laboratory of Alternate Electrical Power System with Renewable Energy Sources, North China Electric Power University, Baoding 071003, China)

Abstract: Traditional transformer fault diagnosis method has poor classification effect in the condition of small example and low recognition efficiency of mass monitoring data. Aiming at the problems above, this paper proposes a parallel variable predictive model based on the Spark computing framework. Firstly, HDFS is used as memory storage system, RowMatrix is stored as a distributed matrix storage structure, and broadcast variable and adjustment of the partition number are used to optimize the degree of parallelism. Secondly, the optimal model type and model parameters are obtained by training four mathematical models to diagnose the transformer fault. The experimental results show that the proposed algorithm has higher accuracy than support vector machine, better computing efficiency than stand-alone environment, and good adaptability to high dimensional feature vector.

This work is supported by National Natural Science Foundation of China (No. 5167702).

Key words: fault diagnosis; small sample; variable predictive model; Spark computing framework; memory storage system

0 引言

随着智能电网建设不断推进，变压器在线监测数据呈现出体量大、类型多、变化快和价值密度低的大数据主要特征。尤其在电网发生故障时，电力设备的监测数据因频繁发送报警数据造成数据井喷，这对监测系统的数据处理能力提出很高的要求，因此通过研究变压器大规模监测数据实现潜伏性故障预测具有重要意义。

实际运行中的电力设备发生故障的几率很低，因此电力设备故障样本十分有限[1]，属于小样本问题。解决电力设备的小样本问题常用算法有支持向量机(SVM)、神经网络、相关向量机(RVM)、集合经验模态(EEMD)及粒子群[2-4]等算法，其发展相对成熟并得到广泛应用。但 SVM 在解决小样本问题时性能受样本本身特性影响较大[5-6]，其中文献[7]将 SVM 应用于图像数据分类，实验结果表明样本数量对其分类性能影响较大，还存在核函数参数和规则化参数选取困难等问题。神经网络虽具备较强的非线性映射能力，但存在局部收敛速度慢和优化时间较长等缺陷[8]。而 RVM 在训练过程会出现矩阵逆奇异及核函数难以确定等问题[9]。

上述传统算法在诊断过程未考虑特征值之间的多变量关联[10]，无法充分利用数据的内在关联。2008 年，相关学者提出变量预测模型分类(Variable Predictive Model Based Class Discriminate, VPMCD)，

基金项目：国家自然科学基金项目资助(5167702)

通过已知样本各特征值间的内在关系建立数学模型，进而预测样本的特征参数，通过最小化特征参数预测误差实现样本多分类，适合处理非线性多分类及小样本问题，已经应用于机械诊断[11-12]、生物识别[18]等领域。因此考虑将该算法应用于电力变压器故障诊断优化问题中。

越来越多大数据处理平台像 Hadoop、Spark 能够分布式并行处理大数据。Hadoop 因频繁读写分布式文件系统成为磁盘 I/O 瓶颈，难以满足实时性需求[14]；而 Spark 具备面向计算任务的实时性，通过将数据分布缓存在节点中，节省大量磁盘 I/O，极大地提升处理速度[15-16]，如文献[17]基于大数据平台进行参数并行化估计，大大提高运行效率。而 Matlab 等传统单机工具面对大量的稠密矩阵计算操作和海量数据显得无能为力，单机环境下训练过程存在数据加载时间和模型训练时间过长以及磁盘读写代价过高等问题，运行效率低下。针对该问题，大数据处理平台给出很好的解决方案，如文献[18]采用云计算平台进行海量数据存储管理，实现快速准确的故障定位，有效解决传统计算工具面对海量数据处理复杂及循环迭代运算困难等问题。

因此考虑结合 Spark 平台构建并行化变量预测模型，并将其应用于海量变压器故障数据的特征提取以及故障分类器的模型学习，解决变压器故障的多分类诊断问题。

1　DGA 诊断基本原理

就电力变压器在线监测而言，存在多种形式状态参数：油中溶解气体、变压器振动信号以及绝缘子泄漏电流[19-20]等。对于油中溶解气体，油中溶解气体分析法(Dissolved Gas Analysis, DGA)通常检测分析变压器油中溶解的几种低分子烃类气体(如氢气 H_2、甲烷 CH_4、乙烷 C_2H_6、乙烯 C_2H_4、乙炔 C_2H_2)等气体浓度、浓度变化趋势以及产生速率等参数[21]，实现变压器内部的潜伏性故障诊断。参考《变压器油中溶解气体分析和判断导则》(以下简称《导则》GB/T7252-2001)，变压器故障类型包括低能放电 D1、高能放电 D2、中低温过热 T12、高温过热 T3 及局部放电 PD 这 5 种故障类型，加上正常情况共构成 6 种类别的故障诊断优化问题。

经典变压器故障诊断分析方法有特征气体法、IEC 三比值法、罗杰斯比值法等[22]。其中 IEC 三比值法通过 C_2H_2/C_2H_4、CH_4/H_2、C_2H_4/C_2H_6 3 个比值大小判断故障类型，因具有应用方便、判断直观以及易于区分局部过热和局部放电的优点得到广泛应用，但存在故障编码不足等问题。《导则》通过更改相关故障类型编码得到改良三比值法，实用性更强，准确率较高。因此本文采用改良三比值法处理数据。

其中改良三比值法编码规则如表 1 所示。

表 1　改良三比值法编码规则

Table 1 Code rule of improved three-ratio method

气体比值范围	C_2H_2/C_2H_4	CH_4/H_2	C_2H_4/C_2H_6
<0.1	0	1	0
≥0.1~<1	1	0	0
≥1~<3	1	2	1
≥3	2	2	2

根据编码组合实现故障类型判断，其中部分故障类型判断方法如表 2 所示。

表 2　故障类型判断方法

Table 2 Judgement method of fault type

C_2H_2/C_2H_4	CH_4/H_2	C_2H_4/C_2H_6	故障类型判断
0	0	1	低温过热
	2	0	低温过热
	2	1	中温过热
	0,1,2	2	高温过热
	1	0	局部放电
1	0,1,2	0,1,2	低能放电
2	0,1,2	0,1,2	高能放电

2　变量预测模型原理

假设采用 $X=[X_1,X_2,\cdots,X_m]$ 表示一种故障类型的特征向量，其中 m 表示特征向量的维数，故障类别中特征值 X_i 受到其他一个或多个特征值的影响。在变量预测模型中，针对特征值 X_i，可以用以下 4 种线性或非线性数学模型 VPM_i 表示，分别为线性(Linear, L)模型、线性交互(Linear Interaction, LI)模型、二次(Quadratic, Q)模型、二次交互(Quadratic Interaction, QI)模型，可表示如下。

1) 线性模型

$$X_i=b_0+\sum_{j=1}^{r}b_jX_j$$

2) 线性交互模型

$$X_i=b_0+\sum_{j=1}^{r}b_jX_j+\sum_{j=1}^{r}\sum_{k=j+1}^{r}b_{jk}X_jX_k$$

3) 二次模型

$$X_i=b_0+\sum_{j=1}^{r}b_jX_j+\sum_{j=1}^{r}b_{jj}X_j^2$$

4) 二次交互模型

$$X_i=b_0+\sum_{j=1}^{r}b_jX_j+\sum_{j=1}^{r}b_{jj}X_j^2+\sum_{j=1}^{r}\sum_{k=j+1}^{r}b_{jk}X_jX_k$$

其中，r(0<r<m)表示模型阶数，且$i \neq j$。在故障诊断过程，通过特征值$X_j(j \neq i)$预测特征值X_i，上述4种模型在预测时均可用式(1)表示。

$$X_i = f(X_j, b_0, b_j, b_{jj}, b_{jk}) + e \tag{1}$$

式(1)为变量X_i的预测模型VPM_i，式中：X_i表示被预测变量；X_j表示预测变量；b_0、b_j、b_{jj}、b_{jk}为模型参数；e表示预测值与实际值的误差。

2.1 VPM 模型建立

假设故障类型为g个，每种故障类型样本数为$n_k(k=1,2,\cdots,g)$，特征向量维数为p，总样本数为n，总样本训练集合可表示为$S[n,p;g]$，VPMCD 训练过程可表示如下。

1) 读取训练集N，按照故障类型将训练集划分为多个互不相容的子集$S_k[n_k,p]$，提取每个样本的特征向量$[X_1,X_2,\cdots,X_m]$，4种数学模型 L、LI、Q、QI 编号为 1、2、3、4。

2) 选取故障类型 k=1，模型类型 m=1，模型阶数 r=1。

3) 针对第 k 类n_k个故障样本集合S_k选取某个样本，对其每个特征值X_i采用模型类型为 m 及阶数为 r 的模型进行建模预测，最终对每个特征值均可得到n_k个 VPM 模型。

4) 采用n_k个 VPM 对模型参数进行预测估计，得到预测值$X_{i\gamma,\text{pred}}(\gamma=1,2,\cdots,n_k)$，利用最小预测误差平方和$\sum_{\gamma}^{n_k}(X_{i\gamma}-X_{i\gamma,\text{pred}})^2$作为判别函数，即可得到特征值$X_i$的预测模型$VPM_i^k$。

5) 改变 k，循环 3)、4)，直至 k=g，得到全部故障类型的 VPM 模型矩阵，如式(2)所示。

$$\begin{bmatrix} VPM_1^1 & VPM_2^1 & \cdots & VPM_p^1 \\ VPM_1^2 & VPM_2^2 & \cdots & VPM_p^2 \\ \vdots & \vdots & & \vdots \\ VPM_1^g & VPM_2^g & \cdots & VPM_p^g \end{bmatrix} \tag{2}$$

6) 改变 m、r 值，循环步骤 3)和 4)。

7) 不断更新 VPM 模型矩阵，最后得出最优预测模型矩阵。

2.2 VPM 模型分类

针对未知故障类型的特征向量$[X_1,X_2,\cdots,X_m]$，预测识别过程如下：

1) 针对特征向量X_i分别采用模型VPM_i^k预测特征值$X_{i,\text{pred}}$，得到 g 个预测特征向量预测矩阵$\left[X_{1,\text{pred}}^k, X_{2,\text{pred}}^k, \cdots, X_{m,\text{pred}}^k\right]$；

2) 采用最小预测误差平方和作为判别函数，分别计算预测向量与 g 个预测矩阵的误差值 e，即$\sum_{t=1}^{p}(X_i - X_{i,\text{pred}})^2$；

3) 将未知故障类型的特征向量归类于误差值 e 最小的类别 k 中，完成故障分类。

2.3 VPM 模型学习

VPM 模型建立过程中关键是求取各模型参数，就式$X_i = f(X_j, b_0, b_j, b_{jj}, b_{jk}) + e$而言，即为$b_0$、$b_j$、$b_{jj}$、$b_{jk}$模型参数的求取，转化矩阵可表示为

$$\boldsymbol{X} = \boldsymbol{\Phi\theta} + \boldsymbol{\xi} = \overline{\boldsymbol{X}} + \boldsymbol{\xi} \tag{3}$$

式中：$\boldsymbol{X}$表示特征向量；$\overline{\boldsymbol{X}}$表示特征向量的预测值；$\boldsymbol{\Phi}$表示设计矩阵；$\boldsymbol{\theta}$表示系数矩阵；$\boldsymbol{\xi}$表示误差向量。误差矩阵可由式(3)转化表示为

$$SSE = \sum_{t=1}^{p}(X_i - \overline{X_i})^2 \tag{4}$$

由此，通过梯度下降法求解 SSE 最小值得到系数向量，完成一次 VPM 模型的学习过程。

综上，采用改良三比值法将气体数据进行编码，对训练样本进行故障类型判别得到数据标签，通过变量预测模型中 4 种数学模型训练带有标签的样本数据，得到模型参数矩阵，对测试样本进行测试，根据误差最小值得到样本故障类型。

3 基于并行 VPMCD 的故障优化

3.1 数据处理

3.1.1 数据输入

就原始数据处理过程而言，在现场采集得到的无标签的油中溶解气体监测数据，经过数据预处理后可以得到带有标签(已知故障类型)的特征气体含量比值，将其作为特征量。

DGA 原始数据由 H_2、H_4、C_2H_6、C_2H_4、C_2H_2 这 5 种气体的含量(μL/L)构成，因原始数据在数量级差异较大，为降低由此造成的计算精度，需要对其进行标准化归一处理，使数据处于 0~1 范围内，处理方式为

$$X_{\text{new}} = \frac{X - X_{\text{mean}}}{X_{\max} - X_{\min}} \tag{5}$$

式中：X 表示原始气体含量；X_{new} 表示标准化后气体含量；X_{mean} 表示数据集中该类气体的平均含量；$X_{\max}$ 表示该类气体中最高含量值；$X_{\min}$ 表示该类气体中最低含量值。

3.1.2 数据并行化

在 Spark 计算框架中，并行化编程有数据并行、任务并行以及混合型并行 3 种方案。其中数据并行

通常采用数据划分实现并行计算，任务并行一般采用任务分解的方式实现并行计算，混合型指结合数据并行和任务并行实现并行化。

而变量预测模型在训练不同故障类型数据时互相独立，对不同故障类型数据进行最小二乘法[23]估计误差前无依赖关系，并且样本数据间处理的先后顺序对实验结果没有影响，具有天然的数据可分性。单机环境下，需要对所有样本顺序依次处理，随着数据规模的增加模型训练效率逐渐降低。因此在集群环境下，通过将不同故障类型的数据并行分散到多个集群节点同时进行处理时，可以减少整体的模型训练时间，大大提高数据的处理效率。

3.2 VPMCD 并行化

在 Spark 计算框架中，并行化程序需要考虑以下几个问题：数据结构的存储方式；如何将程序划分为互相独立的子模块，实现任务的并行化；如何充分发挥 RDD 的缓存优势等实现程序的调优。

3.2.1 分布式存储

随着数据量和特征值维数的增多，VPMCD 模型训练迭代次数大规模增加，最终会生成一个多维 VPM 模型参数矩阵，该矩阵的规模与特征值维数、模型阶数及模型参数正相关。单机环境下内存无法完全存储这些数据，因此采用 HDFS 作为存储系统实现分布式存储数据。

运算过程涉及设计矩阵及矩阵的正三角化分解(QR)等大量稠密矩阵运算，且 Spark 默认按行处理 RDD 数据，因此采用 RowMatrix[24]面向行存储的分布式矩阵存储结构。在逻辑上，RDD 按照 Partition(分区)进行分块实现在集群各节点的分布式，即一个 RDD 内部包含多个并行单元 Partition，通过将并行单元分配到集群中节点实现并行化处理；RDD 以数据块 Block 进行物理存储在内存中。

本文通过 RowMatrix(RDD[Vector])构建实例，其中 RowMatrix 为本地向量的封装类，每行为一个本地矩阵向量 Vector，其中 RowMatrix 在集群中分布式存储结构如图 1 所示。

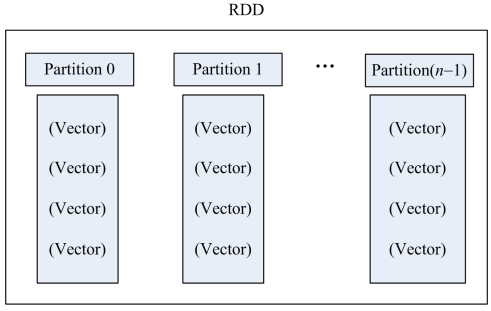

图 1 矩阵的 RDD 存储结构

Fig.1 RDD storage structure of the matrix

3.2.2 并行化结构

并行化 VPMCD 算法涉及主要的数据结构包含原始油中溶解气体数据、变量预测模型 VPM、模型阶数、模型参数以及故障类型等。其中设计类 VPM 数据结构如下：

VPM(RE: Array[Double],modelType: Array[Int],beta: Array[Matrix],flg: Array[Int])

并行 VPMCD 算法过程与单机算法过程类似，其算法伪代码如表 3 所示。

表 3 并行化 VPMCD 算法

Table 3 Parallel VPMCD algorithm

输入：原始数据 D，故障类型数量 g，模型类型 m，模型阶数 r
输出：模型矩阵 VPM

```
SP-VPMCD()
for(i<- 0 to n)
{ broadcast Di                          //广播故障类型 i 的数据矩阵 Di
for(j<- 0 to m)
{   for(k<- 0 to r)
   { var numCom =                       //模型类型与预测变两个数的组合个数
       descartesRDD(d,m)
     var numVars = numOfVars(j,k)       //该种组合下方程的系数个数
     var design=x2fx(d,j)               //将因子设置矩阵转换为设计矩阵
     var QR=design.qr()                 //QR 矩阵
     var RE=Di*QR                       //模型误差
     VPM.set                            //设置 VPM 模型参数
}}
VPM.setRE= min(RE)                      //通过比较模型误差，取得最优模型
(VPM)}                                  //返回模型参数
```

3.2.3 数据状态迁移

油中溶解气体数据集存放在 HDFS 文件系统中，通过 textFile()函数读入文件，创建 RDD 数据切片[25]，经过 collect()函数获取每种故障类的数据总数，对数据集进行 map()、randomSplit()等函数处理数据标签及特征值，并进行数据划分得到训练集和测试集；通过 cartesian()、filter()等函数对数据矩阵进行正三角分解、估算预测值误差等操作，得到每个样本的特征值参数模型；经过 mapPartition()、reduce()等函数进行最小二乘法求取某种故障类型的最优化 VPM 模型；最后将模型参数通过 saveAsTextFile()函数保存在 HDFS 文件系统中，便于模型调用。数据流图如图 2 所示。

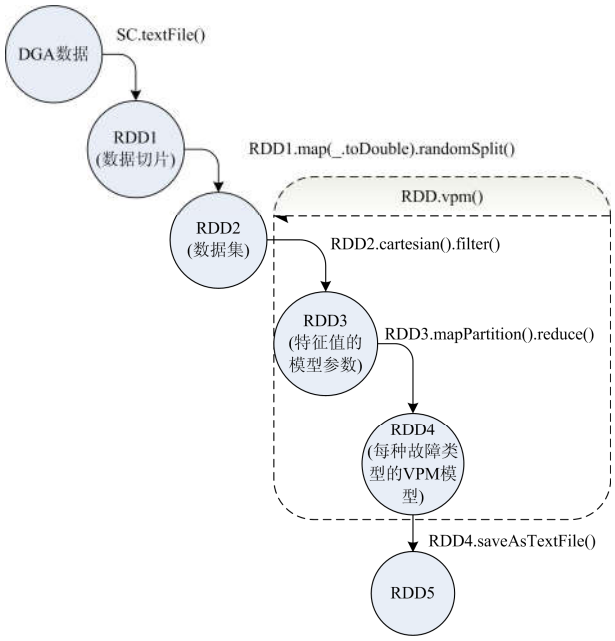

图 2 数据流图

Fig. 2 Data flow diagram

3.2.4 性能调优

本文中主要涉及广播变量及优化并行度两方面。集群环境下，Spark 进行并行化计算时，在各节点间进行数据交换，当数据量较大或者节点数目较多时，数据在节点间迁移造成较大的通信开销。在 Spark 平台，广播(Broadcast)变量为计算过程公用变量保存为节点上只读变量数据，节省存储空间同时也保证数据的传输效率。因此，采用 Broadcast 变量优化节点间数据传输，提升整体的运行速度。此外，Spark 使用高效的广播算法进行变量分发，进一步减少通信开销。采用 SparkContext.broadcast() 方法调用广播变量：

val broadcast tVar = sc.broadcast(v)

此外通过调整合适的分区数减少节点间通信开销。根据集群 CPU 核心数进行调整输入的 RDD 分区数，默认情况下 Spark 为文件的每个 Partition 分配 64 MB 空间，通过 SparkContext.textFile()更改函数传递分区数，或调用 repartition(num)将 RDD 分区为 num 个分区。本文采用这两种方式进行调整分区数。

3.3 基于并行 VPMCD 的变压器故障诊断

基于并行 VPMCD 的变压器故障诊断流程主要包括 3 个部分：

1) 首先构造数据样本集，采集变压器油中溶解气体数据，进行数据标准归一化及改良三比值法处理，构造训练集及测试集；

2) 其次构造并行化变量预测模型，采用带有标签的训练数据集训练变量预测模型的 4 种数学模型，对于每种故障类型通过最优化最小二乘法，得到模型参数进而构造诊断分类模型；

3) 故障诊断：根据训练模型对待测试样本集进行分类，确定故障类型。

其中并行化 VPMCD 故障诊断流程如图 3 所示。

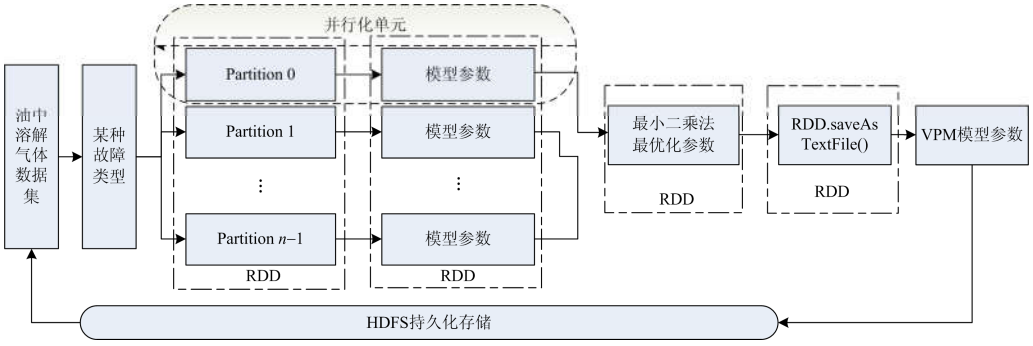

图 3 并行 VPMCD 故障诊断流程

Fig. 3 Fault diagnosis flow of parallel VPMCD

4 实验与仿真证明

4.1 实验环境

Spark 平台搭建中，物理集群为 5 台服务器，其中 1 台作为主服务器(Dell R210II×1)，其余 4 台作为从服务器(ThinkServer RD460×4)，服务器间采用千兆交换机互连，在每台从服务器虚拟出 2 个从节点，共计 1 个主节点(MasterNode)和 8 个从节点(SalveNode)，详细配置如表 4 所示。

4.2 单机环境下分类精度

单机环境下，采用 Matlab 测试该算法性能，运行平台为联想电脑，处理器为 i5-3317，运行内存为 6 G，64 位 Win7 操作系统，编程环境为 Matlab R2014a。训练集为 360 条 DGA 数据，测试集为 483 条测试数据，对 VPMCD 算法和支持向量机进行测试，诊断结果如表 5 所示。

表 4 环境配置

Table 4 Environment configuration

配置项	详细信息
主节点	处理器 i3-2120 3.30 GHz; 内存 3 GB; 硬盘 250 GB
从节点	处理器 E5-2609 v2 2.50 GHz; 内存 7 GB; 硬盘 300 GB
交换机	型号 guidway S1724 G, 传输速率 10/100/1000 Mbps
操作系统	Ubuntu 14.04 LTS
Hadoop	Apache Hadoop-2.6.0
Spark	Apache Spark-1.6.0
JDK	JDK1.8

表 5 故障诊断的精度对比

Table 5 Accuracy comparison of fault diagnosis

类型	SVM/%	VPMCD /%
T12	83.33	88.00
T3	93.33	95.63
D1	90.00	85.71
D2	83.33	90.38
PD	90.00	68.75
正常	85.00	73.33
总计	87.14	88.82

由表 5 可见，VPMCD 算法分类准确率较高，说明 VPMCD 算法具有很好的分类效果，可以满足实际应用需求。此外，在此环境下，VPMCD 算法的训练时间为 1.477 656 s，360 条测试数据的测试时间为 0.350 s，模型训练时间较短。

4.3 集群环境下性能对比

1) 运行时间

在集群环境和 Matlab 单机环境进行模型训练，通过合理的方式对变压器油中溶解气体浓度进行合理扩充，离线 DGA 数据量从 120 条至 2 300 000 条增加，记录模型训练时间如图 4 所示。

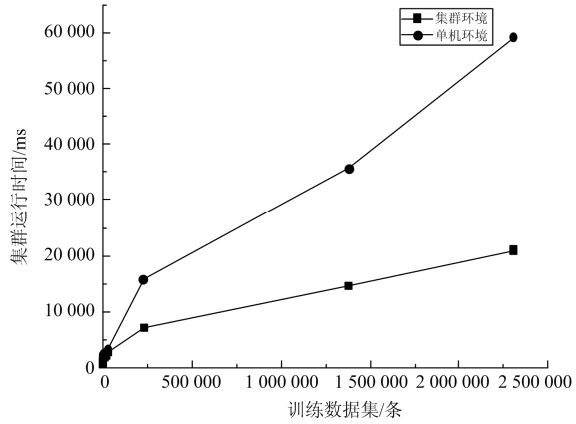

图 4 单机和集群环境下，训练时间与训练数据集的关系

Fig. 4 Relationship between training data set and training time in stand-alone and cluster environment

从图 4 可见，数据规模在百万条时，集群环境下模型训练时间远小于单机环境，并行环境计算显示出优势。

2) 训练得到模型参数

集群环境下，训练样本为 500 条时得到 6 种变压器运行状态下各个预测变量的模型参数如表 6 所示，其中 f 表示模型类型，r 表示模型阶数。

表 6 VPMCD 最佳模型类型及最佳模型阶数

Table 6 Optimal model type and optimal model order of VPMCD

被预测变量	正常状态		低能放电		高能放电		中低温		高温		局部放电	
	f	r	f	r	f	r	f	r	f	r	f	r
X_1	LI	4	QI	3	LI	4	LI	4	LI	4	QI	3
X_2	QI	3	LI	3	QI	3	LI	3	LI	4	QI	4
X_3	LI	3	QI	4	QI	3	Q	3	Q	4	LI	4
X_4	QI	2	LI	3	Q	4	LI	2	QI	3	QI	2

3) 特征向量维数与训练时间

实验过程中发现，随着特征变量维数增加，变量预测模型训练时间急剧增加。为深入探究特征变量维数对变量预测模型的影响，本文采用 UCI 数据集中 Wine 数据集在单机模式和集群模式进行测试分析，针对不同的特征向量维数和训练集样本数进行模型训练。其中该数据集共有 3 个类别，13 维特征，共计 178 个样本。实验结果如图 5 所示。

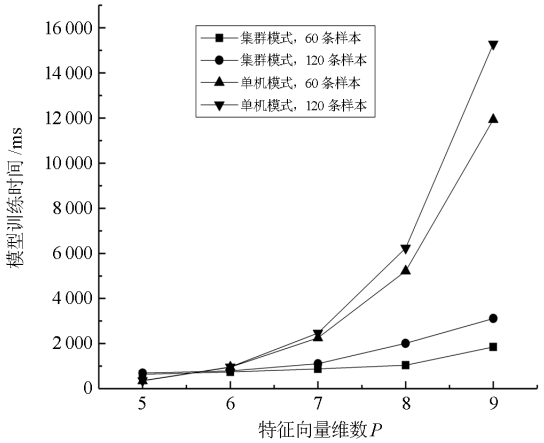

图 5 单机和集群环境下，训练时间与特征向量维数和样本数的关系

Fig. 5 Relationship between training time and the feature vector dimension and the number of samples in stand-alone and cluster environment

从图 5 可以看出，单机模式下随着特征向量维数的增加，模型训练时间成指数型急剧增长；随着训练样本数的增加，特征向量维数对模型训练时间

的影响更加明显。

Spark 集群模式下，在特征维数较小且训练样本较少时，较多时间耗费在集群中节点通信，因此样本训练时间较长；模型训练时间与特征维数和训练样本数为缓慢线性正相关的关系。由此可见，基于 Spark 平台的并行化 VPMCD 算法适合处理多维特征向量的海量数据。

5 结论

本文通过研究变压器海量监测数据下故障诊断分类，提出一种基于 Spark 计算框架的并行变量预测模型。该算法结合 HDFS 文件存储系统，采用改良三比值法处理油中溶解气体数据，将其封装为 RDD，经过一系列转化和行动操作算子实现数据并行化和算法并行化，最后进行实验验证该模型性能。实验结果表明：该模型识别精度高于支持向量机模型，对数据规模、特征向量维数有较好的适应性和扩展性，具有较强的实用价值。

参考文献

[1] 王刘旺, 朱永利, 李莉, 等. 基于特征子集的变压器局部放电小样本类型识别[J]. 电测与仪表, 2015, 52(24): 40-45.
WANG Liuwang, ZHU Yongli, LI Li, et al. Partial discharge type recognition of small sample for power transformer based on feature subsets[J]. Electrical Measurement & Instrumentation, 2015, 52(24): 40-45.

[2] 周云龙, 张岗. 基于集合经验模态分解样本熵和 LIBSVM 的离心风机故障诊断方法[J]. 热力发电, 2017, 46(2): 114-119.
ZHOU Yunlong, ZHANG Gang. A new fault diagnosis method for centrifugal fans based on ensemble empirical mode decomposition sample entropy and LIBSVM[J]. Thermal Power Generation, 2017, 46(2): 114-119.

[3] 田松峰, 胥佳瑞, 王美俊, 等. 基于 EEMD 云模型与 SVM 的汽轮机转子故障诊断方法[J]. 热力发电, 2017, 46(4): 111-114.
TIAN Songfeng, XU Jiarui, WANG Meijun, et al. A rotor fault diagnosis method based on EEMD cloud model and SVM[J]. Thermal Power Generation, 2017, 46(4): 111-114.

[4] 刘沛, 高岳林, 郭伟. 一种基于改进的磷虾群和粒子群的混合算法[J]. 河南师范大学学报(自然科学版), 2017, 45(2): 119-124.
LIU Pei, GAO Yuelin, GUO Wei. A hybrid algorithm based on improved krill herd and particle swarm optimization[J]. Journal of Henan Normal University (Natural Science Edition), 2017, 45(2): 119-124.

[5] XIAO Y, KANG N, HONG Y, et al. Misalignment fault diagnosis of DFWT based on IEMD energy entropy and PSO-SVM[J]. Entropy, 2017, 19(1): 6-12.

[6] 武中利, 杨建, 朱永利, 等. 基于粗糙集理论和支持向量机的变压器故障诊断[J]. 电力系统保护与控制, 2010, 38(18): 80-83.
WU Zhongli, YANG Jian, ZHU Yongli, et al. Power transformer fault diagnosis based on rough set theory and support vector machines[J]. Power System Protection and Control, 2010, 38(18): 80-83.

[7] GHOGGALI N, MELGANI F, BAZI Y. A multiobjective genetic SVM approach for classification problems with limited training samples[J]. IEEE Transactions on Geoscience & Remote Sensing, 2009, 47(6): 1707-1718.

[8] 张颖超, 王雅晨, 邓华, 等. 基于 IAFSA-BPNN 的短期风电功率预测[J]. 电力系统保护与控制, 2017, 45(7): 58-63.
ZHANG Yingchao, WANG Yachen, DENG Hua, et al. IAFSA-BPNN for wind power probabilistic forecasting[J]. Power System Protection and Control, 2017, 45(7): 58-63.

[9] LIU Y, ZHANG J, MA L. A fault diagnosis approach for diesel engines based on self-adaptive WVD, improved FCBF and PECOC-RVM[J]. Neurocomputing, 2016, 177(C): 600-611.

[10] RAGHURAJ R, LAKSHMINARAYANAN S. Variable predictive models - a new multivariate classification approach for pattern recognition applications[J]. Pattern Recognition, 2009, 42(1): 7-16.

[11] LUO S, CHENG J, WEI K. A fault diagnosis model based on LCD-SVD-ANN MIV and VPMCD for rotating machinery[J]. Shock and Vibration, 2016(1): 1-10.

[12] LUO S, CHENG J, ZENG M, et al. An intelligent fault diagnosis model for rotating machinery based on multi-scale higher order singular spectrum analysis and GA-VPMCD[J]. Measurement, 2016, 87(1): 38-50.

[13] RAYMOND W, ILLIAS H, ABU B A, et al. Partial discharge classifications: review of recent progress[J]. Measurement, 2015, 68(1): 164-181.

[14] WHITE T. Hadoop: the definitive guide[M]. O'Reilly Media, Inc. 2012.

[15] ZAHARIA M, CHOWDHURY M, FRANKLIN M J, et al. Spark: cluster computing with working sets[C] // Usenix Conference on Hot Topics in Cloud Computing, USENIX Association, May 2010, Berkeley, USA: 10-10.

[16] 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 Jianliang, LIU Dechao. 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.

[17] 何迪, 章禹, 郭创新, 等. 一种面向风险评估的输电线路故障概率模型[J]. 电力系统保护与控制, 2017, 45(7): 69-76.
HE Di, ZHANG Yu, GUO Chuangxin, et al. Failure probability model of transmission lines for risk assessment[J]. Power System Protection and Control, 2017, 45(7): 69-76.

[18] DENG Feng, ZENG Xiangjun, PAN Lanlan. Research on multi-terminal traveling wave fault location method in complicated networks based on cloud computing platform[J]. Protection and Control of Modern Power Systems, 2017, 2(2): 199-210. DOI: 10.1186/s41601-017-0042-4.

[19] 郑一鸣, 何文林, 孙翔, 等. 基于油色谱超立方映射的电力变压器缺陷援例诊断模型[J]. 电力工程技术, 2017, 36(4): 48-53.
ZHENG Yiming, HE Wenlin, SUN Xiang, et al. Case based power transformer defeats diagnose model using hypercube mapping of oil chromatography[J]. Eectric Power Engineering Technology, 2017, 36(4): 48-53.

[20] 陆云才, 胡汉巧, 蔚超, 等. 基于超声波法的变压器重症监护系统研制及应用[J]. 电力工程技术, 2017, 36(2): 94-98.
LUN Yuncai, HU Hanqiao, WEI Chao, et al. Development and application of transformer intensive care system based on ultrasonic[J]. Eectric Power Engineering Technology, 2017, 36(2): 94-98.

[21] 陈欢, 彭辉, 舒乃秋, 等. 基于鲁棒能量模型 LS-TSVM 和 DGA 的变压器故障诊断[J]. 电力系统保护与控制, 2017, 45(21): 134-139.
CHEN Huan, PENG Hui, SHU Naiqiu, et al. Fault diagnosis of power transformer based on RELS-TSVM and DGA[J]. Power System Protection and Control, 2017, 45(21): 134-139.

[22] 公茂法, 张言攀, 柳岩妮, 等. 基于 BP 网络算法优化模糊 Petri 网的电力变压器故障诊断[J]. 电力系统保护与控制, 2015, 43(3): 113-117.
GONG Maofa, ZHANG Yanpan, LIU Yanni, et al. Fault diagnosis of power transformers based on back propagation algorithm evolving fuzzy Petri nets[J]. Power System Protection and Control, 2015, 43(3): 113-117.

[23] LARSSON E, SHCHERBAKOV V, HERYUDONO A. A least squares radial basis function partition of unity method for solving PDEs[J]. Siam Journal on Scientific Computing, 2017, 39(6): 1-24.

[24] 卞昊穹, 陈跃国, 杜小勇, 等. Spark 上的等值连接优化[J]. 华东师范大学学报(自然科学版), 2014(5): 263-270.
BIAN Haoqiong, CHEN Yueguo, DU Xiaoyong, et al. Equi-join optimization on spark[J]. Journal of East China Normal University (Natural Science), 2014(5): 263-270.

[25] KUNJIR M, FAIN B, MUNAGALA K, et al. ROBUS: fair cache allocation for multi-tenant data-parallel workloads[C] // ACM International Conference on Management of Data, May 14-19, 2017, Chicago, USA: 219-234.

收稿日期：2018-04-11；　修回日期：2018-05-29

作者简介：

马利洁(1994—)，女，硕士研究生，研究方向为电力设备大数据分析；E-mail: malijie_spring@qq.com

朱永利(1963—)，男，通信作者，教授，博士生导师，主要研究方向为网络化监控与智能信息处理研究；E-mail: yonglipw@163.com

郑艳艳(1987—)，女，博士研究生，主要研究方向为电力设备在线监测与故障诊断。

(编辑　周金梅)