计算机科学
COMPUTER SCIENCE

一种基于SCD文件的合并单元高速数据压缩方法

陈星田, 熊小伏, 白勇, 胡海洋

引用本文

陈星田, 熊小伏, 白勇, 胡海洋. 一种基于SCD文件的合并单元高速数据压缩方法[J]. 计算机科学, 2023, 50(12): 123-129.

CHEN Xingtian, XIONG Xiaofu, BAI Yong, HU Haiyang. High Speed Data Compression Method of Merge Unit Based on SCD File [J]. Computer Science, 2023, 50(12): 123-129.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230700230

一种基于SCD文件的合并单元高速数据压缩方法

陈星田[1]　熊小伏[2]　白　勇[1]　胡海洋[2]
1 重庆电力高等专科学校　重庆 400053
2 输配电装备及系统安全与新技术国家重点实验室(重庆大学)　重庆 400044

摘　要　在现代智能电网中,智能变电站安装了大量合并单元来同步发布电流互感器和电压互感器的暂态量,这些暂态数据有必要保存长达数年,从而覆盖设备生命周期,为设备状态维修、可靠性等研究提供原始信息支撑,但是如此长时与高频的海量数据给存储设备带来了巨大压力。文中首先将高频暂态数据分为固定不变的、状态变化的和周期变化的3种形式来进行预处理,将固定不变部分用SCD文件中的唯一标识代替,状态变化部分用事件记录文件代替,周期变化部分用SCD文件中双通道差量和周期差量来表示。然后使用16位哈夫曼完成最终压缩编码,并对比测试了各种预处理前后的压缩结果和不同编码的压缩结果。最终的测试结果表明该压缩方法比普通硬件压缩卡压缩比更大,压缩速率比普通压缩卡更快。

关键词:合并单元采样值;无损数据压缩;哈夫曼编码;LZMA压缩算法;小波变换
中图法分类号　TP391;TM734

High Speed Data Compression Method of Merge Unit Based on SCD File

CHEN Xingtian[1],XIONG Xiaofu[2],BAI Yong[1] and HU Haiyang[2]
1 Chongqing Electric Power College,Chongqing 400053,China
2 State Key Laboratory of Power Transmission Equipment & System Security and New Technology(ChongQing University),Chongqing 400044,China

Abstract　In modern smart grid,many merging units are installed in smart substation to release transient data of current transformer and voltage transformer synchronously,these transient data need to be saved for several years,so as to cover the life cycle of equipment and provide original information support for condition maintenance and reliability of equipment,but such long-time and high-frequency massive data is a difficult problem for storage equipment. In this paper,the high-frequency transient data are preprocessed in three forms:fixed,state-changing and periodic-changing. Tthe fixed part is replaced by merge's APPID in SCD file,the state-changing part is replaced by event record file,and the periodic-changing part is represented by two-channel difference and periodic difference in SCD file,and the final compression coding is completed with 16-bit Huffman. The final test shows that the compression ratio of this compression method is larger than that of common hardware compression card,and the compression rate is faster than that of common compression card.

Keywords　Merge unit sampling data,Lossless data compression,Huffman coding,LZMA compression algorithm,Wavelet transform

1　引言

当前国内智能变电站模拟量采样主要采用合并单元MU(Merge Unit)来同步多路电流以及电压瞬时数据,然后按照国际电工委员会IEC61850-9-2标准规定的ISO/IEC8802-3[1]格式发给保护、测控和计量等设备。这些数字化后的模拟量是一种近似正弦或余弦波在不同时刻的瞬时采样,是无限循环波形的离散数据值。由于智能变电站模拟量采样点通道较多,每秒采样次数也较多,因此MU所产生的数据量巨大[2]。而输变电设备的状态维护必须依赖于长期的设备数据分析,才能对设备的可靠性进行评估与故障预测[3-7]。相关文献显示,互感器的平均无故障时间约为62352h[8],即7.1年,而网络报文记录及分析装置只要求数据记录的时间大于3天[9],这么短的数据记录时间,无法全部覆盖互感器的全寿命时间。目前智能变电站网络报文记录及分析装置为了尽可能增加存储数据并降低中央处理器(CPU)的负荷,一方面扩展磁盘空间;另一方面采用硬压缩卡等技术对数据进行压缩[10],压缩方法采用RFC1952(Gzip)方法,但是这种压缩卡压缩比并不

到稿日期:2023-07-31　返修日期:2023-10-31
基金项目:重庆市自然科学基金(CSTB2022NSCQ-MSX0251)
This work was supported by the Natural Science Foundation of Chongqing,China(CSTB2022NSCQ-MSX0251).
通信作者:陈星田(30160954@qq.com)

高。笔者购得国外某公司的一张硬压缩卡，官方宣称压缩比在6∶1左右，但这种压缩比只能在文本数据压缩中才能达到，在实际工作发现，该卡针对MU数据平均压缩比只有2.4∶1左右，压缩速率每秒在31.40 MB～43.69 MB之间。在220 kV～750 kV电压等级规模的智能变电站中，每秒产生的数据量平均在80 MB～120 MB之间，特殊的变电站达到170 MB，如表1所列，需要配置多块压缩卡才能满足数据处理要求，机器中功耗很容易超出相关技术规范的100 W的要求[9]，而且这种低压缩比的数据在存储和传输时还会带来空间和带宽紧张等一系列后续问题。数据压缩的需求与数据产生的速度高度相关，高速大数据应尽可能在数据产生的相同时间内压缩完毕，否则难以应用在数据产生期。考虑到目前智能变电站的数据规模，有必要找到一种新压缩算法，其压缩比需大于6∶1的硬压缩卡压缩比，压缩速率每秒需提升到80 MB～120 MB左右才能具有较好的适用性。

表1 智能变电站MU每秒数据统计

Table 1 Smart substation MU per second data statistics

序号	变电站MU数目	数据帧数目	全部数据帧大小/字节
1	750 kV达坂城(138)	552000	156456000
2	750 kV西山变(95)	380000	106000000
3	500 kV南阳中(87)	348000	88040000
4	500 kV广丰变(35)	140000	41508000
5	220 kV大朗变(160)	640000	169700000
6	220 kV金山变(54)	216000	57704000

2 当前主要数据压缩方向相关工作

国内外关于数据压缩的研究很多，根据最终还原的结果来看，数据压缩可以分为有损压缩和无损压缩，有损压缩主要应用于视频和图像，即使某些数据丢失，对用户的感官也没有太大的影响，代表算法有基于离散余弦变换的数据压缩、基于小波的数据压缩和基于线性预测编码的数据压缩等，不适用于智能变电站这种需要数据能精确还原的场景。无损压缩主要有Huffman、LZ77[11]、游程编码、区间编码(LZMA)等[12]，最终演变成基于统计的字典编码和基于区间的概率编码。还有许多方法是在这种两种编码的基础上衍生出的数据预处理方法[13-14]。还有一些方法是将字典压缩Gzip算法和LZMA迁移至现场可编程门阵列(FPGA)[15-16]进行并行处理，进而提提高数据压缩处理速度，但是这些成果未转化为实用商品出现在市场上。

基于统计的压缩方法最典型的代表就是Huffman编码[17]，该方法是根据数据信息中符号重复出现的概率生成的一种前缀编码方法，它和香农-范诺(Shannon-Fano)编码[18]一样，核心都是构造二叉树，只是Huffman是在扫描符号出现概率排序后，自最低频率向最高频率构造二叉树。静态Huffman编码压缩结果需要保存构造二叉树以方便解压数据时再构造二叉树，但8位的二叉树比较小，每个符号只需要保存符号频率和符号共5字节到压缩文件。如图1所示，256个符号加符号出现的频率，最多用1 280个字节就可以存储，Huffman编码的压缩速度非常快，耗时和文件大小线性相关，但是当数据文件符号出现概率没有明显差别时，它的压缩比不高。

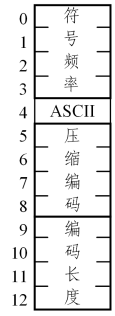

图1 8位哈夫曼符号结构体

Fig. 1 8bit Huffman symbol structure

另一种基于统计的压缩方法是LZMA，它是7Zip采用的压缩方法。LZMA是在LZ77[11]算法的基础上加入了基于比特流的区间编码以及基于动态规划的深度优化，并扩大了数据压缩字典(DEFLATE)，最大到4GB[19]，压缩效果十分接近数据的熵值。但是随着数据和字典的增大，LZMA压缩方法的压缩时间会迅速增加，对于智能变电站采样值而言，这种数据压缩的时间已经超过了数据生成的时间。

国内还有针对数据采集与监视控制系统SCADA(Supervisory Control And Data Acquisition)数据[20]和录波数据的压缩方法[21-22]，以线性拟合、FFT变换以及小波变换对数据进行预处理，再利用Huffman或LZMA等方法进行最终的编码压缩，取得了较好的压缩效果。虽然这种变换主要是针对浮点数据，也是有损压缩，但是该方法很容易应用到整型数据和无损数据压缩中[23]。

目前国内所有数据压缩方向的研究在编码理论方面都没有太大突破，基本上都是沿用Huffman和LZMA来完成最终编码，但是它们针对不同类型的数据进行不同的预处理，最终的压缩结果差异很大。文献[24]提出的基于图像差分和神经网络压缩算法就是一种数据预处理方法，图像差分和本文提出的双AD通道差量原理相同，神经网络是一种更新的数据预测方法，与LZMA区间的预测目标一致，都是要得到数据概率分布，进而排序算术编码，达到压缩的目的。但是该方法需要GPU硬件支持才能完成，其功耗大部分都大于100 W，目前还不适合嵌入式应用。本文基于变电站配置文件SCD(Substation Configuration Description)的MU高速数据压缩方法在提出一种全新的数据预处理方法的同时，也改进了常规Huffman编码长度，是在数据预处理和最终编码方面都有改进的数据压缩方法，在压缩比、压缩速率和功耗方面具有一定优势。

3 基于SCD配置文件的采样值处理模型

对于智能变电站任何一个MU采样数据帧，都可以用一个有限长度字符序列x_t来表示，如式(1)所示：

$$x_t = x_0 x_1 \cdots x_k \tag{1}$$

其中，t表示某一时刻；k为某一MU数据长度，单位为字节，其值取决于SCD对于该MU的所配置关联的通道数，即互感器数目。

x_t在不同的时间是不同的，但是并不是所有的数据都完全不同，因此可以把x_t分解为 3 个部分来表示，如式(2)所示：

$$x_t = w_c + e_t + \vec{v}_t \tag{2}$$

其中，w_c表示该部分数据在任何时刻都是不变的常量序列，这个常量序列可以根据合单元的应用标识(Application Identification，APPID)从 SCD 配置文件中间接或直接获得；e_t表示变化很小的状态量序列，这种变量并不总是在变化中，而且变化后也会保持一段时间；$\vec{v}_t$表示多路模拟量的采样序列$(v_{A1}, v_{A2}, v_{B1}, v_{B2}, v_{C1}, v_{C2}, \cdots)$，该项表示变电站某点的电压或电流的瞬时值，基本上一直随时间变化。在一段时间内采集到此 MU 所有的数据构成的文件，用连续报文集合 A 来表示，如式(3)所示，则未压缩前集合 A 可以表示为多个w_c的集合、多个e_t集合与多个$\vec{v}_t$集合的总和。需要指出的是，w_c，e_t，$\vec{v}_t$在一帧数据中并不连续，e_t序列中分散有w_c项，$\vec{v}_t$序列中也分散有e_t项，这是根据数据封装规约来定义的。

$$A = \sum_{t_1}^{t_n} x_t = (t_n - t_1) w_c + \sum_{t_1}^{t_n} e_t + \sum_{t_1}^{t_n} \vec{v}_t \tag{3}$$

如果用$z_t = F(z_t, x_t)$来表示压缩处理过程，集合 A 通过z_t处理后，结果集合可表示为A'，如式(4)所示：

$$A' = F(z_t, x_t) \tag{4}$$

z_t针对 3 部分数据分别进行预处理，即：

$$A' = F(z_t, w_c) + F(z_t, e_t) + F(z_t, \vec{v}_t) \tag{5}$$

w_c是常量数据序列，可全部看成是冗余数据，因此z_t在预处理时直接将其筛除，$F(z_t, w_c)$直接定义为 0，如式(6)所示：

$$F(z_t, w_c) = 0 \tag{6}$$

针对变化量小的数据e_t，$F(z_t, e_t)$预处理时生成一个事件文件来记录状态量的变化，当数据变化时，记下该时刻与状态量值，如果没有变化，就不用记录。因此 $F(z_t, e_t)$的预处理有两种结果，如式(7)所示：

$$F(z_t, e_t) = \begin{cases} 0, & e_t = e_{t-1} \\ e_t, & e_t \neq e_{t-1} \end{cases} \tag{7}$$

$\vec{v}_t$可以被看成是多路模拟采样信号，在 110 kV 及以上智能变电站 MU 设计中，为了提高可靠性，采用双模数转换(AD)形式为保护提供二路相同的电压和电流，由于相同二路 AD 数据误差较小，因此也有较大冗余。在保存一路 AD 值后，另一路保存差值即可，差值的值域远小于原值，甚至很多时候都是零值，这正是压缩算法所希望达到的目标。因此，$F(z_t, \vec{v}_t)$第一步是进行通道间差量处理，将双 AD 采样之间进行差量预处理，如式(8)所示：

$$\begin{aligned} F(z_t, \vec{v}_t)_1 &= \overrightarrow{v_t'} \\ &= (v_{A1}, v_{A2} - v_{A1}, v_{B1}, v_{B2} - v_{B1}, v_{C1}, v_{C2} - \\ &\quad v_{C1}, \cdots) \end{aligned} \tag{8}$$

此外，由于$\vec{v}_t$是交流电的电压电流采样信号，因此可以将其看成是多(通道)路近似正弦信号组合。每个通道是由 50Hz 的基波和多个高次谐波组成，如式(9)所示：

$$\begin{aligned} v_{tx} &= a_0 + A_0 \sin(\omega_0 t + \varphi_0) + A_1 \sin(\omega_1 t + \varphi_1) + \\ &\quad A_2 \sin(\omega_2 t + \varphi_2) + \cdots + \Delta_x \end{aligned} \tag{9}$$

其中，$\omega_k = 2k\pi\omega_0$，Δ_x是各谐波浮动量之和。

$F(z_t, \vec{v}_t)$第二步就是利用周期函数特性进行周期间差量预处理，具体方法为：用当前值减去一个周期前的值，保存值域更小的结果。v_{tx}表示当前时间的某通道值，$v_{(t-T)x}$表示一个周波前的某通道的值，$\overrightarrow{\Delta_x'}$表示多个通道与前周波的差值序列。

$$F(z_t, \overrightarrow{v_t'})_2 = \overrightarrow{v_t''} = \overrightarrow{v_{tx} - v_{(t-T)x}} = \overrightarrow{\Delta_x'} \tag{10}$$

结合式(6)、式(7)、式(9)、式(10)完成对数据的预处理，如式(11)所示：

$$A' = \sum_{t_1}^{t_n} F(z_t, e_t) + \sum_{t_1}^{t_n} \overrightarrow{\Delta_x'} \tag{11}$$

3.1 采样值数据帧结构与属性

根据 IEC61850-9-2 标准[1]，智能变电站采样值以太网封装结构如图 2 所示，数据帧捕获 PCAP(Packet Capture)头部由捕获以太网数据时的硬件生成，包括捕获的时标和数据长度标识，接着是以太网 ETH(Ether)头部和虚拟网络标识 VLAN(Visual Lan)，然后是包含 APPID 和长度的采样值(Sampling Value，SAV)头部。APPID 是一个十六位的数据值，是 SCD 文件配置时分配给 MU 应用的唯一标识，长度占 2 个字节。最后是抽象语法标记 ASN. 1(Abstract Syntax Notation dotone)封装的采样值应用规约数据单元(Sampling Value Protocol Data Unit，SAVPDU)，SAVPDU 里面封装了每个通道采样值的应用服务数据单元(Application Service Data Unit，ASDU)。

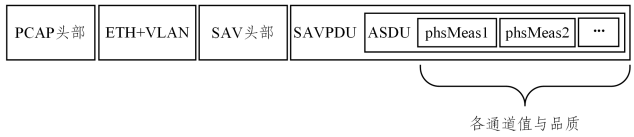

图 2 采样值帧结构

Fig. 2 Sample value frame structure

采样值 ASDU 数据内容是由 SCD 配置文件中的 IED 配置给 MU 的采样值发送数据集(Dataset)决定，数据集由多个数据条目组成，如图 3 所示。每个数据条目由一个 32 位有符号整数和一个 32 位无符号整数组成，如图 4 所示，32 位有符号整数表示电压或电流的瞬时值(Analogue Value)，可以看成是正弦或余弦曲线上的离散值。其中，如果条目是电流，则比原始值扩大 1 000 倍，如果条目是电压，则比原始值扩大 100 倍。32 位无符号整数表示 32 位有符号整数的品质的状态量(Quality)，只用了低 14 位，高 18 位均无效。

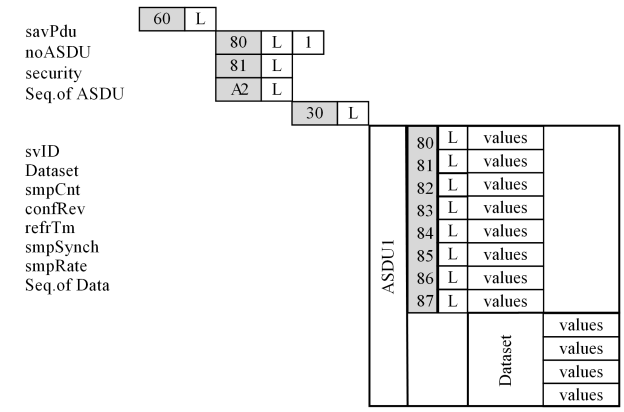

图 3 SAVPDU 的 ASN. 1 封装结构

Fig. 3 ASN. 1 package structure of SAVPDU

MU 数据集都会有一个条目表示 MU 额定延时，即该采样从模拟量转换为数量到发布出来的延时，如表 2 所列。额定延时条目在硬件没有变化时发送的内容是固定不变的。无符号整数表示条目的数据品质，数据品质在不同的数据采集时期可能会有变化。

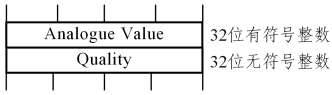

图 4 每个通道数据条目结构

Fig. 4 Structure of each channel data entry

表 2 某主变低压侧 MU 数据集

Table 2 Dataset of merge unit at low-voltage side of a main transformer

序号	数据对象	描述
1	MUSV/LLN0. DelayTRtg	MU 额定延时
2	MUSV/SVOUTTCTR1. Amp	保护 A 相电流 Ia1
3	MUSV/SVOUTTCTR1. AmpR	保护 A 相电流 Ia2
4	MUSV/SVOUTTCTR2. Amp	保护 B 相电流 Ib1
5	MUSV/SVOUTTCTR2. AmpR	保护 B 相电流 Ib2
6	MUSV/SVOUTTCTR3. Amp	保护 C 相电流 Ic1
7	MUSV/SVOUTTCTR3. AmpR	保护 C 相电流 Ic2
8	MUSV/SVOUTTCTR4. Amp	计量 A 相电流 Ia
9	MUSV/SVOUTTCTR5. Amp	计量 B 相电流 Ib
10	MUSV/SVOUTTCTR6. Amp	计量 C 相电流 Ic
11	MUSV/SVOUTTVTR1. Vol	保护 A 相电压 Ua1
12	MUSV/SVOUTTVTR1. VolR	保护 A 相电压 Ua2
13	MUSV/SVOUTTVTR2. Vol	保护 B 相电压 Ub1
14	MUSV/SVOUTTVTR2. VolR	保护 B 相电压 Ub2
14	MUSV/SVOUTTVTR3. Vol	保护 C 相电压 Uc1
16	MUSV/SVOUTTVTR3. VolR	保护 C 相电压 Uc2
17	MUSV/SVOUTTVTR4. Vol	计量 A 相电压 Ua
18	MUSV/SVOUTTVTR5. Vol	计量 B 相电压 Ub
19	MUSV/SVOUTTVTR6. Vol	计量 C 相电压 Uc
20	MUSV/SVOUTTVTR7. Vol	零序电压 3U01
21	MUSV/SVOUTTVTR7. VolR	零序电压 3U02
22	MUSV/SVOUTTVTR8. Vol	同期电压 Ux1
23	MUSV/SVOUTTVTR8. VolR	同期电压 Ux2

根据《110 kV～220 kV 智能变电站设计规范》以及《330 kV～7500 kV 智能变电站设计规范》，保护用电压准确度不低于 3P(3%)，保护用电流准确度不低于 5 TPE(一次电流是额定一次电流的 10 倍时，绕组的复合误差≤±5%)。MU 对继电保护发布采样数据时，电流和电压都采用双 AD 采样，这两个 AD 的类型、精度和变比等硬件都相同，因此两通道的数据在保护误差范围内是基本一致的，保护装置任取一个都可。

连同 MU 数据集一同封装入 ASDU 的，还有采样同步标志、配置版本、计数器序号以及 svID(Sampled Values Identification)，封装的规约是 ASN. 1。同步标志表明 MU 是否接收到时钟同步脉冲，计数器每次采样增加 1，其他项在 SCD 配置完成后都是固定不变的，如图 3 所示。ASDU 和 noASDU(ASDU 数目)项封装在一起，构成 SAVPDU 采样值应用数据单元。然后与 APPID 采样值应用 ID 封装在一起构成以太网数据帧，如图 5 所示。VLANID 和 VLAN 优先级在配置 VLAN 时一起封装在以太网数据帧里给交换机标明优先级处理顺序。除了数据帧源 MAC 地址外，目的地址 MAC 地址、VLANID、VLAN 优先级、APPID、svID、配置版本等内容在 SCD 配置文件中都有设置。

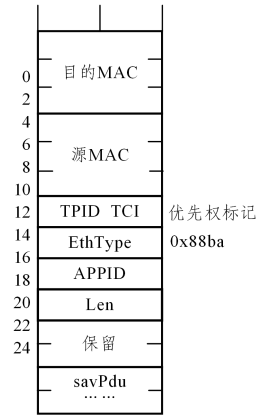

图 5 采样值 ETH 帧结构

Fig. 5 Sample value ethernet frame structure

在以太网数据采集后，使用 PCAP 数据结构帧头来记录采集帧的时间和捕获长度，如图 6 所示。时间分为两部分，都是 32 位无符号整数，tv_sec 表示距 1970 年 1 月 1 日零时零分零秒的秒数，tv_usec 是微秒数，另外还有两个 32 位整数分别表示捕获数据的长度和帧的本来长度，一般情况下两者大小一致，只有在捕获的数据不完整时数据长度和捕获长度才不相等，这种情况要归属到异常事件中。

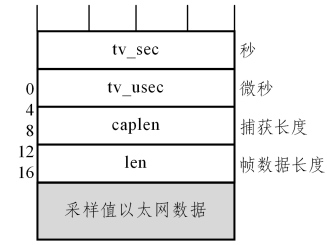

图 6 PCAP 帧头结构

Fig. 6 PCAP frame head structure

3.2 采样值数据预处理

通过上述采样值分析可以发现，很多数据在采样值发送时基本上是固定不变的，比如 MAC 地址、VLANID 和 VLAN 优先级等，只需要用 APPID 唯一标识即可，预处理第一步就是建立一个 MU 配置文件，将这些不变的数据项 w_c 部分放到 MU 的配置文件中，另外创建一个事件记录表，将状态变化量 e_t 项的变化用一个新的事件记录方式记入事件记录表文件中，从而在每帧报文中去掉这些冗余数据。数据还原时，通过 APPID 找到 MU 的配置信息，并扫描事件记录表，还原所有的变化事件。最终简化后的帧结构如图 7 所示，所有报文头部只剩下 16 个字节。新报文头部中，APPID 继续沿用表示应用标识，长度用来表示新的数据长度，smpCnt，Sync，sec，usec 等继续用来表示采样计数器、同步标志、采样时间的秒、采样时间的微秒等。紧随头部的是所有通道的 32 位采样值，采样品质等状态量都移入事件记录文件中，事件记录文件中每条记录使用一样的报文头部。

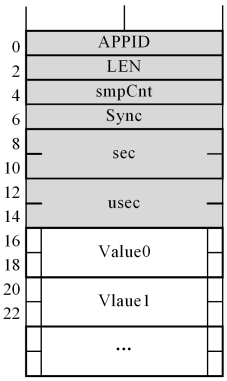

图 7 简化后采样值帧结构

Fig. 7 Simplified sample value frame structure

此外，为了提高后面编码的效率，应尽可能产生较多的重复数据，针对正弦或余弦这种波形数据，可利用傅里叶变换或整数小波变换这种数据预处理将曲线数据分解为多次谐波幅值与相角量，其值域会大幅缩小，但是傅里叶变换需要进行大量浮点计算，消耗很多 CPU 时间。整数小波变换对一个序列数据需要多次分组求差，也会消耗较多 CPU 时间，而且整数小波变换在处理重复的数据时，会达到反向效果，故不采用。下面介绍不需要大量数值计算的数据预处理方法。

如前所述，智能变电站 IED 采样值控制块所用的数据集中，保护电流和电压一般采用双 AD 形式，$F(z_t,\vec{v_t})_1$ 预处理设计用第二个 AD 通道中的数据量减去第一个 AD 通道中的数据量，将这个差量保存在第二个通道中，就可以产生很多相同的差量。根据表 2 所列的数据集条目，可以设计 8 组双 AD 差量组合，如表 3 所列。

表 3 双 AD 差量组组合

Table 3 Double-AD differential group combination

组合序号	描述
1	保护 A 相电流 Ia1
	保护 A 相电流 Ia2
2	保护 B 相电流 Ib1
	保护 B 相电流 Ib2
3	保护 C 相电流 Ic1
	保护 C 相电流 Ic2
4	保护 A 相电压 Ua1
	保护 A 相电压 Ua2
5	保护 B 相电压 Ub1
	保护 B 相电压 Ub2
6	保护 C 相电压 Uc1
	保护 C 相电压 Uc2
7	零序电压 3U01
	零序电压 3U02
8	同期电压 Ux1
	同期电压 Ux2

由于不同的 MU 各个条目不尽相同，双 AD 差量组合方式要在数据预处理前读取，因此需要提前将 SCD 文件解析，将双 AD 差量组合方式保存在配置文件中。

考虑到采样数据大部分都是基波频率在 50 Hz 左右的正弦波和各种整数次的谐波组合，在非故障期，相邻两个周期之间的数据是比较接近的，因此 $F(z_t,\overrightarrow{v_t'})_2$ 预处理定义为用后一个周波的数据减去前一个周波的数据，将此差量保存在第二个周期中，以此类推，除了第一个周波外，可以产生很多相同的差量。

$$\overrightarrow{v_t''}=\overrightarrow{v_{tx}-v_{(t-T)x}}=\overrightarrow{\Delta_x'}$$
$$=v_{c,k}-v_{c,(k-N)\%N} \tag{11}$$

其中，c 表示第 c 个通道（条目）；k 表示当前采样序号，实际是计数器 smpCnt 值；N 为周波采样点数，$(K-N)\%N$ 表示前一个周波的索引值。

4 压缩编码处理

目前广泛采用的编码的方式为 LZMA，但 LZMA 不是一个稳定的编码方式，它需要花很多时间来寻找滑动窗口中的匹配字符串，这个时间并不确定，只有在时间足够的情况下，采用此方法才可获得较大压缩比。智能变电站中 MU 较多，数据处理时间非常有限，因此必须采用稳定迅捷的编码方法。

Huffman 编码是一种很稳定的编码方法，花费的时间与数据大小完全成正比。传统的 Huffman 采用 8 位编码，频度字典只有 256 个字节，只要扫描一遍就可以统计出各个字节的频率，然后按频率高低进行编码，频率最高的编码最短，最低的编码最长。即使使用最少编码完成的压缩，最大压缩比也不会超过 8:1，这已经是 8 位 Huffman 压缩的极限了。

由于现在广泛使用 64 位操作系统，因此可采用 16 位 Huffman 编码迅速提高压缩比，改进后的编码结构体设计如图 8 所示，其与图 1 的主要区别在于 Ascii 长度使用 16 位，可显著提高压缩比。

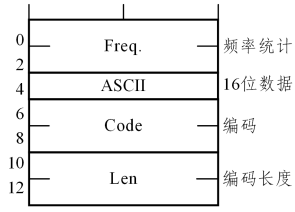

图 8 16 Huffman 编码结构体

Fig. 8 16 Huffman coding structure

5 压缩测试

从某 750kV 变电站获得真实 MU 暂态数据，分别截取 100 MB，200 MB，300 MB 进行压缩测试，并完成 SCD 文件解析与 MU 信息以及双 AD 差量组合配置的生成准备工作。3 个暂态数据文件的字节数、帧数和生成持续时间如表 4 所列。

表 4 待测试暂态数据帧数与时间

Table 4 Transient data frames and time for test

文件字节数	帧数	时间/s
104 857 806	320 666	80.166 249
209 715 261	641 331	160.332 499
314 573 370	961 998	240.499 249

5.1 测试 $F(z_t,w_c)$ 和 $F(z_t,e_t)$ 预处理效果

让 3 种不同大小的文件通过 $F(z_t,w_c)$ 和 $F(z_t,e_t)$ 预处理，剔除可以从 SCD 文件中生成、且在每帧报文中固定不变的数据，以及状态量的冗余数据，最后处理结果如表 5 所列。由于 $F(z_t,e_t)$ 处理数据时本段数据没有出现状态量变化，

因此事件记录文件大小为零。最后通过 $F(z_t,w_c)$ 和 $F(z_t,e_t)$ 预处理后，结果数据大小迅速减小 50% 以上，效果非常显著。

表 5 $F(z_t,w_c)$ 预处理效果

Table 5 $F(z_t,w_c)$ pretreatment effect

原始文件/MB	$F(z_t,w_c)$ 预处理后/MB	$F(z_t,w_c)$ 预处理耗时/ms
100	41.336	1283.559
200	82.672	2354.348
300	124.008	3543.654

5.2 测试 $F(z_t,\vec{v}_t)$ 预处理效果

将表 5 中的 3 个文件进行 $F(z_t,\vec{v}_t)$ 预处理，再进行 16 位 Huffman 编码，由于 $F(z_t,\vec{v}_t)$ 预处理并不影响数据大小，因此只是修改数据值域。为了比较 $F(z_t,\vec{v}_t)$ 预处理效果，表 5 中的 3 个文件不进行 $F(z_t,\vec{v}_t)$ 预处理，直接进行 16 位 Huffman 编码，两者结果对比如表 6 所列。从表中可以看出，经过 $F(z_t,\vec{v}_t)$ 预处理的数据编码比未经 $F(z_t,\vec{v}_t)$ 预处理的数据小 50% 以上，可见 $F(z_t,\vec{v}_t)$ 预处理对压缩有显著影响。

表 6 $F(z_t,\vec{v}_t)$ 预处理效果

Table 6 $F(z_t,\vec{v}_t)$ pretreatment effect

(单位:MB)

原始文件	$F(z_t,w_c)$ 后直接编码	$F(z_t,w_c)+F(z_t,\vec{v}_t)$ 后编码
100	26.336	12.544
200	52.392	25.371
300	78.473	38.066

5.3 基于 SCD 文件压缩与 LZMA 的压缩比较

将 100 MB，200 MB，300 MB 这 3 个文件，通过本文所述的全部预处理后再用 16 位 Huffman 进行编码，另用 LZMA2201 版对 3 个文件进行最大压缩比压缩，最终结果如表 7 所列，从压缩比上来看，LZMA 压缩比较高，可达11:1，但耗时非常长，100 MB 数据压缩时间接近 1 min。而采用 SCD 配置文件预处理并利用 16 位 Huffman 编码后，压缩比可达 8:1。虽然此压缩比小于 LZMA，但其耗时只有 3s 左右，远少于 LZMA 的耗时，算上所有数据的预处理时间，也不及 LZMA 所用时间的 1/10。300 M 文件只用了 LZMA 1/20 的时间就处理完毕。

表 7 基于 SCD 压缩与直接 LZMA 压缩效果比较

Table 7 Comparison between SCD-based compression and direct LZMA compression

原始文件/MB	直接 LZMA 压缩		基于 SCD 压缩	
	耗时/ms	大小/MB	耗时/ms	大小/MB
100	54673.866	8.803	3418.438	12.544
200	104868.408	17.519	5141.896	25.371
300	154103.127	26.279	7645.877	38.066

5.4 比较 16 位和 8 位 Huffman 编码效果

在 $F(z_t,w_c)+F(z_t,\vec{v}_t)$ 预处理的基础上分别使用 16 位和 8 位 Huffman 编码对数据进行编码压缩，对比结果如表 8 所列。

表 8 16 位和 8 位 Huffman 压缩比较

Table 8 Comparison of 16 bit and 8 bit Huffman coding

原始文件/MB	8 位 Huffman 编码		16 位 Huffman 编码	
	耗时/ms	大小/MB	耗时/ms	大小/MB
100	3014.250	16.258	3418.438	12.544
200	6079.060	33.147	5141.896	25.371
300	8439.394	49.836	7645.877	38.066

从表 8 可以看出，16 位 Huffman 和 8 位 Huffman 数据压缩所用时间接近，16 位 Huffman 编码略快于 8 位 Huffman 编码，但 16 位 Huffman 编码压缩比明显优于 8 位 Huffman 编码，且文件越大，效果越明显。

以上测试计时是在台式电脑 i7-6700 双核 4 线程 3.4 GHz CPU、8 GB 内存、Windows 10 操作系统平台下完成，编译器是 Visual C++ 2017。引言中所述硬件压缩卡计时是在目前常用的网络报文记录装置嵌入式电脑 Celeron(R)-2980U 双核 2 线程 1.6GHz CPU、4 GB 内存、CentOS 7.5.1804 操作系统平台下完成的，编译器是 GCC 4.8.5 20150623。所有测试程序计时过程均为单线程计时，此装置可以运行 4 个独立线程，因此可以计算出其综合处理数据的速度在每秒 117.03 MB～156.97 MB 之间。

结束语 从上述压缩测试结果可以看出，基于 SCD 配置文件的数据预处理压缩方法用相同的硬件，最终数据压缩比达到 8:1，最高数据处理速率每秒达到 160MB 左右，单台设备就能满足普通智能变电站全站数据压缩存储要求。此方法主要是利用 SCD 配置文件去除了一些固定不变的冗余数据，把状态量变化数据迁移到事件记录文件，并且利用周期数据的相似性的特点，用差量代替原始，降低了待压缩数据的值域，此方法和傅里叶分解或小波变换的目的一样。基于 SCD 配置文件的数据压缩方法在压缩比上不是最大的，但是压缩处理耗时非常少，压缩速度既快又线性稳定，非常适合用于智能变电站 MU 这种海量高速暂态数据的压缩。由于它避免了复杂的浮点运算，因此适合在 FPGA 中以 Verilog 语言编程实现，为并行采集与压缩处理各种规模的智能变电站暂态大数据提供了一种新算法。同时这种将待压缩数据分解为固定不变、状态量变化和周期性变化这 3 种数据，并分别用不同的预处理手段进行数据预处理的方法，同样可以用在其他高速采样数据压缩中，如高频行波数据和普通录波数据等。

参 考 文 献

[1] DL/T860.92-2006. Communication networks and systems in substations-Part 9-2: Specific Communication Service Mapping (SCSM)-Sampled values over ISO/IEC 8802-3 [S]. Beijing: National Electric Power System Management and Information Exchange Standardization Technical Committee, 2006.

[2] FU G X, DAI C J. Integrated design and implementation of network analysis and fault recording for intelligent substation [J]. Electric Power Automation Equipment, 2013, 33(5): 163-167.

[3] HOU A J, XIONG X F, SHEN Z J, et al. A reliability decision method of CBM maintenance schedule for transmission equipment [J]. Power System Protection and Control, 2012, 40(22): 108-112.

[4] LI L, XIONG W, LU D M, et al. Study on the prediction method

for failure rate in the reliability evaluation of power transmission and transformation facility [J]. Electrical Measurement & Instrumentation,2015,52(3):37-41.

[5] TIAN L,XING J G. Discussion on making decision about electric equipment for condition based maintenance [J]. Power System Technology,2004,28(16):60-63.

[6] ZHAO M M,LIN S S,LI Q,et al. Reliability analysis of smart substation secondary equipment [J]. Process Automation Instrumentation,2022,43(4):45-50.

[7] WU X,WANG Y,YIN X G,et al. Development and application of power equipment status evaluation system [J]. High Voltage Apparatus,2020,56(6):7-12.

[8] CUI C,WANG M M,XU Y L,et al. Reliability Prediction of Electronic Current Transformer Based on Rogowski Coil [J]. Shandong Electric Power,2016,43(6):14-17,36.

[9] Q/GDW 10715-2016,Technical Specification for Network Message Recording and Analysis Device of Intelligent Substation [S]. Beijing:State Grid Corporation of China,2017.

[10] WANG X A,DOU Z S,JIN H R,et al. The new realization of messages recorder and analyzer used in smart substation [J]. Electrical Technology,2014(2):82-85.

[11] ZIV J,LEMPLE A. A universal algorithm for sequential data compression [J]. IEEE Transactions on Information Theory,1977,23(3):337-343.

[12] ZIV J,LEMPLE A. Compression of individual sequences via variable-rate coding [J]. IEEE Transactions on Information Theory,1978,24(5):530-536.

[13] XI W,LI P,LI P,et al. Atwo-stage PMU data compression method for edge computing devices of distribution networks [J]. Power System Technology,2023,47(8):3184-3193.

[14] YUE Q M,YU W Y,BAI C J,et al. Novel compression scheme of fault recording data in power systems based on lifting algorithm [J]. Automation of Electric Power System,2005,29(5):74-78.

[15] LI B,ZHANG,LIU Y. FPGA hardware implementation of the LZMA compression algorithm [J]. Journal of Beijing University of Aeronautics and Astronautics,2015,41(3):375-382.

[16] CHEN X J,LI B,ZHOU Q L. Implementation of RTL Scalable High-Performance Data Compression Method [J]. Acta Electonica Sinica,2022,50(7):1548-1557.

[17] HUFFMAN D A. A method for the construction of minimum-redundancy codes [J]. Proceedings of the IRE,1952,40(9):1098-1101.

[18] SHANNON C E. A mathematical theory of communication [J]. The Bell System Technical Journal,1948,27(3):379-423.

[19] Lempel-Ziv-Markov chain algorithm [EB/OL]. https://infogalactic.com/info/Lempel-Ziv-Markov_chain_algorithm.

[20] MA F Y,LI Q P,MA Z B,et al. The Research of Historical Data Compression and Storage Strategy in Power Dispatch SCADA System [J]. Power System Technology,2014,38(4):1109-1114.

[21] HUANG C,YANG S X,LIANG Y C,et al. Practical data compression method for power system fault records [J]. Electric Power Automation Equipment,2014,34(6):162-167.

[22] FEI M W,YUE Q M,ZHANG P C,et al. Wavelets Selection of Compression and Reconstruction Algorithm Based on Digital Recorded Data from a Faulted Power System [J]. Automation of Electric Power Systems,2005(17):64-67,97.

[23] HUANG T S,WANG Y,WU D,et al. Second generation wavelet-based data compression algorithm for power system fault recorder [J]. Electric Power Automation Equipment,2004(3):59-62.

[24] FU S Y,WANG L,CHENG Y D,et al. Synchrotron radiation source image compression method based on difference and neural network [J]. Journal of National University of Defense Technology,2022,44(5):53-62.

CHEN Xingtian, born in 1971, Ph.D, engineer. His main research interest is smart grid automation technology.

(责任编辑:何杨)