DATA MODELING FOR REDUCTION OF VOLUME IN LARGE ARCHIVES OF POWER QUALITY DATA

Jan KRAUS
KMB systems, TUL – CZ
jan.kraus@tul.cz

Viktor BUBLA
TUL – CZ
viktor.bubla@tul.cz

Leoš KUKAČKA
TUL - CZ
leos.kukacka@tul.cz

ABSTRACT

The article presents various options and its experimental and practical results for compression of archives of power quality readings – the main aggregated data from a typical survey. It uses data modeling as a preprocessing step which improves the total compression ratio. Time series of real values of typical quantities from a commercial power quality analyzer are compressed using different algorithms. The data is optionally preprocessed using different models to utilize the semi-predictable features of each time series. As will be shown this technique can effectively improve the compression by several percent points.

INTRODUCTION

One of the typical problems of aggregated and sample data collection in power quality monitoring campaign in a real application is to develop an effective storage and data analysis schema. The same issue arises even more with the growing market penetration of smart grid and smart metering appliances as these will produce enormous volumes of data. The lossy or loseless compression is a natural choice for optimization of these archives. But its actual performance hugely depends on several aspects such as selected quantity coding, ordering of the data, compression block size etc.

In our paper we discuss several loseless techniques for the compression layer. We compare performance of the latest available implementation of different available algorithms. On the data preprocessing layer we make use of the specific aspects of the recorded data and we propose various techniques to reduce the entropy of such data prior to its actual compression. These data preprocessing techniques include mostly modeling and prediction to further improve the achieved compression ratio.

For three phase oscillogram data (voltage and current) it is typical to use the original channel data as well as its various linear combinations to test the optimal set of real and virtual output channels. The modeling technique is one which we refer to as physical models and in a limited fashion it can also be used for the aggregated data.

Usually the compression can be improved due to the signals periodicity. The outputs are picked according to achieved results and in combination that still can be used for the restoration of all original channels. Similar techniques are long being used for image and sound compression (jpeg, mp3). For the compression of disturbance data the application of different transforms is among other described in [7,8].

Auto-correlation can be observed for longer time intervals (such as day-to-day or for shifts periodicity) and many appliances have periodic operation – HVAC cycles etc. So heuristic reordering of the aggregated data or a sort of modeling based on Fourier transform might also prove to be useful. But better technique for the aggregated data proposed by us is to use curve fitting which reduces the entropy of archived dataset with a small amount of additional model data.

We have previously used a virtual instrument to experiment with compression ratios and to pick the optimal combination of encoding and compression. In this article all the experiments are performed on actual datasets of real power quality analyzers (SMV, SMPQ and SIMON) used on different measurement campaigns and for various periods of time.

The outcome of the described experiments and techniques is a proposal of compression method and data format specifically designed for optimal storage of the three phases power monitoring data. It is shown that precious selection of the reduction data model leads to significant savings in the storage space, while providing an useful tool for faster and more advanced data mining in huge measurement archives.

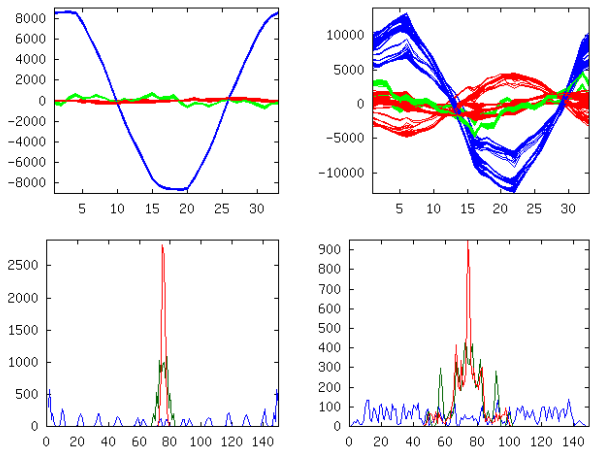

Fig.1: periods of sampled signal (V left, I right) displayed in time-collected view. Histograms of original, phase-to-phase differential and period-to-period differential.

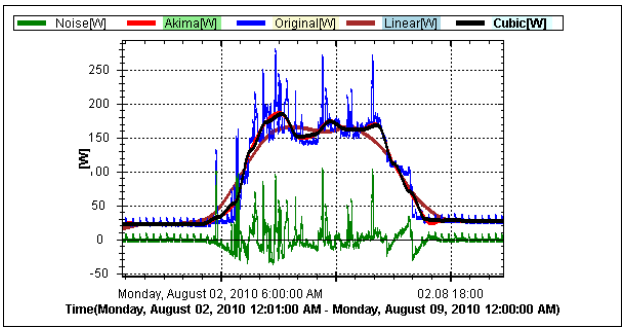

Fig. 2: example of the original data, its various models and the residual noise (of the Akima spline model).

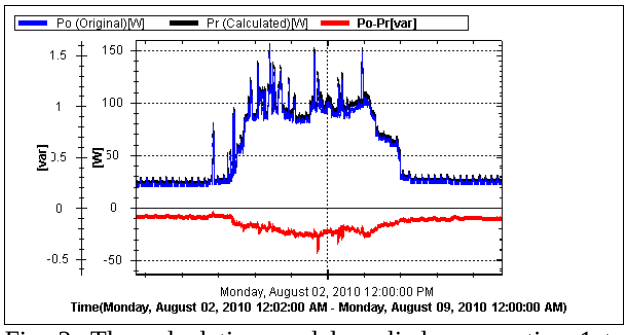

Fig. 2: The calculating model applied on equation 1 to evaluate active power P. The residual noise here is magnified by factor of 100 to reveal details.

MODELS AND COMPRESSION

Models in Respect to Fitness and Compression

We have evaluated each models capability to describe the different time series in data and to reduce the entropy of the noise (difference of the original samples to the model). Entropy was evaluated by the means of average value and its standard deviation. Our expectation is that the limited spread of the 'noise' data would improve the overall compression ratio.

Compression Algorithms

In our previous experiments [3],[4] we have tested compression performance of several public and available compression libraries – from the simpler entropy coders (Huffman, arithmetic, range encoding) through the well known dictionary based algorithms based on the Lempel-Ziv algorithms (ZIP, GZIP) up to the most recent algorithms based on the Burrows-Wheler transformation (BZIP2).
Our interest was to discover different algorithm for small embedded systems and for pc hosted computationally powerful database server. For the later application the BZIP2 algorithm outperformed the other in compression ratios in most cases and was recommended for general use. In the following experiments we have compared ZIP and BZIP2 coders again with the newer LZMA coder. The newly benchmarked LZMA[2] algorithm used in the 7-zip tool is known to perform better than BZIP2 and ZIP. LZMA is based on the LZ77 with advanced modeling optimizations and internally it also builds up a dictionary. It also employs range encoding in its second stage which is an efficient entropy coder. This is its great advantage for our approach where several models are used to reduce the entropy of the tested datasets.

Proposed File Format

None of the tested techniques is the always optimal one. For real application the various datasets can give different result with different models and compressions. Also different intervals in the same data can compress well with another algorithms. For this reasons some compression applications (such as FLAC[5]) use formats with block structure.

Using this principle each block of the data can be encoded, modeled and compressed with a different technique. The compression algorithm than usually tries various approaches to store the same subset of data and chooses the most optimal one out of many. Size of the block can be fixed or variable to improve the previous. Decompression simply uses a magic number in each block to restore the original data.

Modeling of time series in power quality campaign

Averaging Models

AM is a probably one of the simplest aggregation technique. For the application in compression, averaging can be used amongst the time series intervals to approximate the real values with its averages. For the optimal creation of other models mentioned bellow the sliding averaging also is an important tool. It is usually computationally impossible to test all possible selections of the modeling point subsets. Random selection of points can hit local extremes which seriously distracts its actual fitness. The averaging step smooth the input data prior to selection of the subset for models thus providing better results with the uniform distribution of model points.

Polynomial Models

PoM is a traditional mean of modeling for time series. They use set of points and fit a polynomial onto that set while minimizing an error. Its unfortunate disadvantage is an oscillating behavior when modeling complex shapes. Its behavior inside one interval is also influenced by the data in all other intervals. For long time series they are also impractical as they need to store larger number of approximation data to fit the original well.

Spline Models

SM approximates the values in time series with a specific curve for each interval. It can be simple (y=Ax+B) line as well as more complex spline with various other qualities such as fluent continuation at the point of join. Different spline model do provide simple and useful tool for time series modeling. As can be seen most of the used spline based models have performed similarly in regard to the model quality (tab. 2).

Fourier Models
FM is particularly useful for describing periodically occurring behavior. This model is useful for describing any type of seasonality in the data. Its strongest application is to 'filter out' insignificant frequency components of the data.
Physical (Formulae) Models
PhM is a different tool to model set of quantities. Instead of the time series they model the vector of multiple values in a given time with a set of equations that are known to be valid for that original. This can create multiple modeled variables which correspond tightly to its original. A good example of this principle is for example the following relation:

$$(U*I)^2 \approx P^2+Q^2 \quad (1)$$

This equation is valid for each actual reading but in a specific application to aggregated measurement data set it is not exactly correct. To use it for entropy reduction in compression though it can be expected that it would estimate value of one selected variable reasonably well so that the residual difference will be normally distributed with smaller deviation and with center closer to zero.

	U	*I*	*P*	*Q*	*THDI*	*THDU*
LZMA	0.10	0.19	0.40	0.36	0.13	0.10
BZ2	0.08	0.21	0.45	0.43	0.11	0.10
ZIP	0.15	0.30	0.49	0.47	0.20	0.16
LZMA	0.14	0.11	0.27	0.17	0.12	0.09
BZ2	0.13	0.11	0.22	0.16	0.12	0.10
ZIP	0.20	0.16	0.31	0.23	0.19	0.19

Tab. 1: Compression ratios for raw (white) and Akima spline modeled (grey) time series for a week long measured office data with minute aggregation KM1.

EXPERIMENTS AND RESULTS

Several dataset of different sizes and aggregation intervals were used to evaluate performance of compression with and without modeling reduction in place. The single phase readings of voltage, current, power (active, reactive) and voltage and current THD were evaluated separately. Achieved ratios (for dataset KM1, Akima spline) are shown in tab. 1.
These experiments have revealed that PoM's are not very suitable to model the data well enough due to its oscillatory behavior. The averaging and spline models have performed equally well (tab. 2) and for the used data to be approximated well enough it was only required to use roughly 2% of the original data points. This is an important quality as the model (defined by those points) is itself stored in the output stream and thus it adds to the inefficiency of compression.
The best modeling performance for entropy reduction was achieved with the PhM as shown on fig. 2, where the noise is by two orders less than the original signal. Unfortunately this excellent modeling capability is limited by the number of available or applicable relations between quantities in each different data set and thus it limits its overall impact on the total compression ratio.

	AVG	*STD*	*MIN*	*MAX*	*MED*
INP	80.0	91.8	19.2	541.3	31.1
AKI	80.4	85.2	11.2	343.1	30.7
CUB	80.4	84.8	11.9	340.2	30.7
AVG	80.1	81.0	10.8	334.7	31.1
LIN	80.4	84.6	11.9	340.3	30.7
~AKI	1.5	26.7	-123	231	-0.7

Tab. 2: Parameters of sample distribution of original readings of power, its various models and the residual (difference) values (grey) in KM1 data set.

The FM on contrary would require keeping majority of volume of the information as compared to the original data to model it reasonably well. For this reason we did not use it in the final compression ratio comparisons on the main archive datasets. This fact does not limit its applicability on the dataset for other analytical purposes. Also its capability to describe periodicity might be suitable for filtering of the data prior to the modeling stage and also for compression of the raw oscillogram data.
As can be seen the achieved compression ratio varies among different quantities. As expected the worst compression result is given by the ZIP type of compression. LZMA typically performed the best on the original data sets. What is a surprising fact is that after the data are reduced by one of the suitable models, the performance of BZIP2 and LZMA usually becomes much closer.
Often the modeling improves compression of the data significantly (such as for I, P and Q in our case). On the other hand the current and voltage THD are not much influenced by these additional steps. We believe that this happens because the THD values actually recorded are very little distributed in relative comparison which eliminates the positive effect of modeling.
On the case of voltage in our presented result it can be actually seen that modeling can even worsen the ratio. Again this might be for the fact that such a quantity already has distribution that can't be significantly improved by further modeling. For that reason the proposed file structure enables to leave the modeling step out in such cases.

	LZMA			*BZip2*			*ZIP*		
	Or	Mo	No	Or	Mo	No	Or	Mo	No
1	.16	.16	.04	.18	.18	.03	.24	.24	.05
2	.16	.16	.03	.17	.17	.03	.24	.24	.05
4	.18	.19	.06	.21	.22	.06	.27	.27	.09
8	.20	.20	.07	.21	.21	.06	.29	.29	.11

Tab. 3: The original, model and noise data for P calculated according to eq. 1 for four different data sets.

CONCLUSIONS

As was shown the various modeling techniques applied on power quality measurement data can be used to reduce its entropy and thus improve the achievable compression ratio. In the evaluated cases the achieved extra space saving was typically rating between 5 to 15% in comparing with the plain compression algorithms. Optimal models based on curve approximation have been constructed from as little as 2% of the number of measurements of the original datasets. As an additional advantage for the proposed compression method many of the stored models can be used for fast preload/preview of the stored data in visualization software (in a similar manner as jpeg images are used for preview on slow internet connection). Also they are very useful in analytic probation and data mining in large database sets as they mathematically describe its generalized features. Each model can be as well though to describe typical profile of the measured quantities and a way to predict its future behavior or parameters.

Acknowledgments

Authors would like to express appreciation for all support received by the engineers of KMB systems, s.r.o while developing the application data layer as well as for all the instruments and its data used in this experiments.
This work is co-financed from the student grant SGS 2010 - Interactive Mechatronics Systems Using the Cybernetics Principles

REFERENCES

[1] D. Salomon, G. Motta, 2010, *Handbook of Data Compression*, Springer-Verlag, London, 411-416.

[2] I. Pavlov, 2010, Homepage of 7-ZIP/LZMA, *http://www.7-zip.org*

[3] J. Kraus and T. Tobiska and V. Bubla, 2009, Looseless encodings and compression algorithms applied on power quality datasets, *20th International Conference and Exhibition on Electricity Distribution*, 987, ISBN: 978 1 84919 126 5.

[4] J. Kraus and V. Bubla, 2009, Optimal Methods for Data Storage in Performance Measuring and Monitoring Devices, *Sborník konference EPE 2008*.

[5] J. Coalson, 2010, Introduction to FLAC Format, *http://flac.sourceforge.net*

[6] C. Ruegg, 2010, Math.NET project, *http://www.mathdotnet.com*

[7] O. Gerek and D. Ece, 2008, Compression of power quality event data using 2D representation, Electric Power Systems Research, Volume 78, Issue 6, Pages 1047-1052

[8] S. Santoso and E.J. Powers and W.M. Grady, Power quality disturbance data compression using wavelet transform methods, *IEEE Transactions on power Delivery, Volume 12, Issue 3*, 1250-1257, ISSN:0885-8977