HBC: Combining Lossy and Lossless Hybrid Bilayer Compression Framework on Time-Series Data

Wanying Lu* , Liang Liu* , Wenbin Zhai* , Haoyuan Chen* ， Yulei Liu*

*** Nanjing University of Aeronautics and Astronautics**

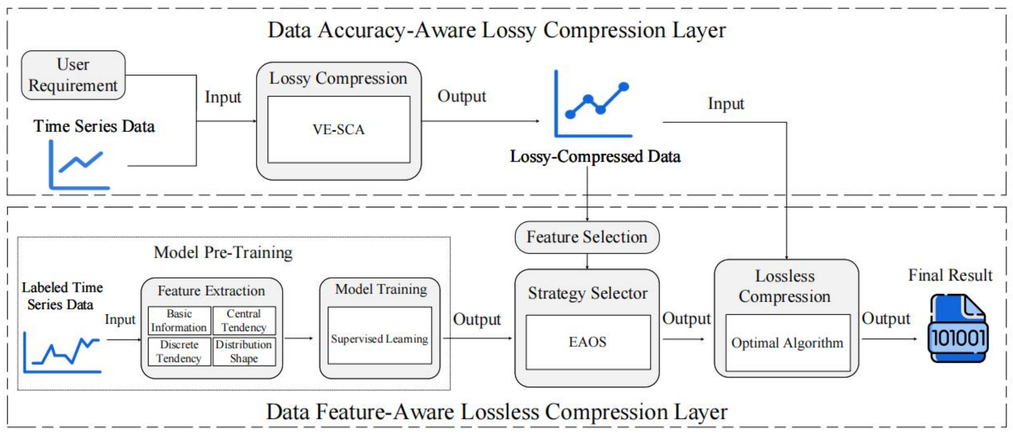
MOTIVATION

Data compression is a key technology for time series data storage, but currently there are still the following challenges in this field.

- Traditional data compression strategies are not universal, and they cannot achieve optimal compression on datasets with different characteristics.
- In practical applications, users have different accuracy requirements for time series data in different numerical ranges. However, traditional data compression strategies implement undifferentiated compression with a uniform error threshold for all data.
- Current data compression strategies cannot achieve high compression ratios and low compression costs at the same time.

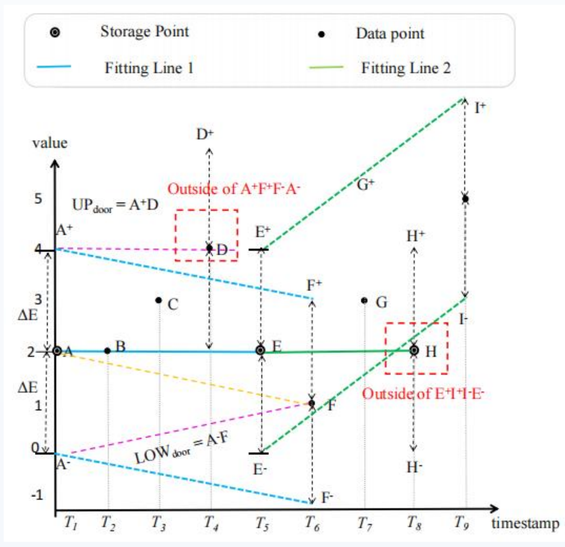
COMPRESSION FRAMEWORK — HBC

To address the current issues of low universality, low compression ratio, and high overhead in data compression, we design a Hybrid Bilayer Compression (HBC) framework. As shown in below figure, HBC includes a data accuracy-aware lossy compression layer and a data feature aware lossless compression layer. The framework adaptively adjusts the error value according to different data ranges to achieve efficient lossy compression, and further selects the optimal lossless compression strategy for the lossy compressed result.



DATA ACCURACY-AWARE LOSSY COMPRESSION LAYER

At the top layer of the framework, we design a lightweight lossy compression algorithm called VE-SCA. It achieves data accuracy-aware lossy compression through piecewise fitting strategy and dynamic error adjustment strategy.



Segment Fitting Strategy: The segment fitting strategies mainly use a linear function to represent a set of continuous data whose data fluctuation is less than the error threshold. We mainly implement a segment fitting strategy by constructing parallelogram. The parallelogram is constructed by four points whose upper and lower sides are ΔE away from the current data point and the previous stored data point, respectively. If the constructed parallelogram can completely contain all the points before the current point, then the fitting of this segment can be continued.

Dynamic Error Adjustment Strategy: We adjust the data error from two aspects. FIirst, we dynamically adjust the error threshold ΔE based on the different numerical ranges. Then, for the same numerical range, we fine-tune the error threshold ΔE based on the data fluctuation.

1) Error Adjustment for Different Numerical Ranges: According to the numerical ranges, we divide the data into several levels of accuracy: $ACC_1, ACC_2, ACC_3, \cdots, ACC_i$. Then, we use different error threshold standards for different levels. The above specific rules can be represented as Equation (4).

2) Error Adjustment for The Same Numerical Ranges: We further consider using data fluctuations to adjust the ACC_i error threshold between $\Delta Emin_i$ and $\Delta Emax_i$. The data fluctuation of time series data, denoted as K, can be represented as Equation (5), where FD is the fitting degree of the previous compression segment and FD' is the other one. The so-called fitting degree refers to how many points can be fitted in a line segment fitting. Then, we use Equation (6) to slightly adjust ΔE within the error threshold for the corresponding accuracy level. Note that in Equation (6)， α is the acceptable threshold specified by the user, $F(k) = (K-1)^3 + 1$.

$$\Delta E \in \begin{cases} [\Delta E_{\min_1}, \Delta E_{\max_1}] & \text{if } D \in ACC_1 \\ [\Delta E_{\min_2}, \Delta E_{\max_2}] & \text{if } D \in ACC_2 \\ \cdots & \\ [\Delta E_{\min_i}, \Delta E_{\max_i}] & \text{if } D \in ACC_i \end{cases} \tag{4}$$

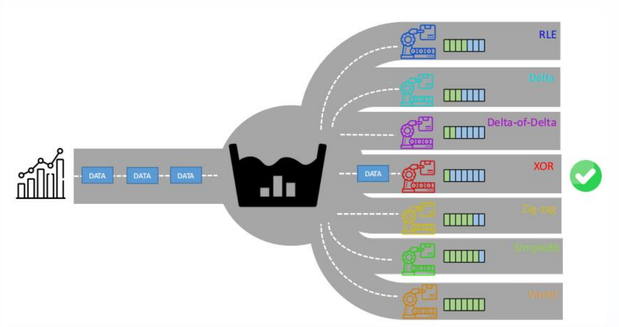
$$K = \frac{FD}{FD'} \tag{5}$$

$$\Delta E_n = \begin{cases} \Delta E_{n-1} F(K) & \text{if } |k-1| > \alpha \\ \Delta E_{n-1} & \text{if } |k-1| \le \alpha \end{cases} \tag{6}$$

DATA FEATURE-AWARE LOSSLESS COMPRESSION LAYER

Data feature-aware lossless compression layer includes an Efficient Adaptive Offline Selector (EAOS) based on supervised learning, which can select the optimal lossless compression algorithm from the compression algorithm pool for D'. EAOS transforms the compression problem into a multi-classification problem of time series data. We formalize a piece of data into a set of data features with a label to obtain a training sample z, which can be expressed as $z = (x, y)$, where x represents the feature vector of the data, namely $x = (BIF, CT F, DT F, DSF)$, and y is the classification label of x: we use “0” ~ “6” to represent the above seven compression methods, and the label is the final selected optimal algorithm. After a period of data sampling, we obtain the labelled training dataset, which can be used with supervised learning to train our EAOS.

Dimensions	Feature	Description
Basic information features	D_{Type}	The data type
	L_{Sign}	The sign bit of the data
	V_{Max}	The maximum values of the data
	V_{Min}	The minimum values of the data
Central tendency features	D_{Mean}	The mean of the data
	D_{Median}	The median of the data
	D_{Mode}	The mode of the data
Dispersion tendency features	σ	The standard deviation of the data
	IQR	The interquartile range of the data
Distribution shape features	D_{Kurt}	The kurtosis coefficient of the data
	D_{Skew}	The skewness coefficient of the data



RESULITS

In this section, we compare HBC with three state-of-the-art compression schemes to demonstrate its efficiency and universality. Meanwhile, ablation experiments are conducted to analyze the performance of the lossy compression algorithm VE-SCA at the upper layer and the lossless compression model EAOS at the lower layer.

Experiment 1： We compared the prediction performance of different supervised learning models. It can find that the MLP model are better than other models, and it can achieve a precision rate of 80% under 20% false positives rate.

Experiment 2： We implement six compression schemes on one-dimensional data. The effect of HBC is the best among all compression schemes. As shown in Fig. 5, its average CR on six one-dimensional datasets is only 1.96%.

Experiment 3： We evaluate the efficiency of six compression schemes. Among HBC, AMMMO, Chimp128, and LFZip, the compression efficiency of HBC is the best, and it can save 15%, 90% and 97% of the CT respectively.

Experiment 4： We evaluate compression errors of HBC and LFZip on six one-dimensional datasets. The compression error of HBC is greater than that of LFZip on most datasets. This is mainly because HBC adopts a dynamic error strategy, while LFZip adopts a unique minimum error strategy.

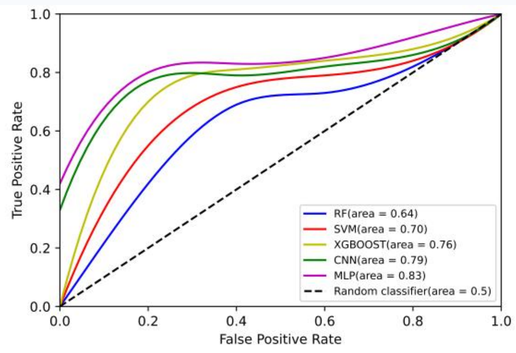


Fig. 4: Comparison of ROC indicators of different models.

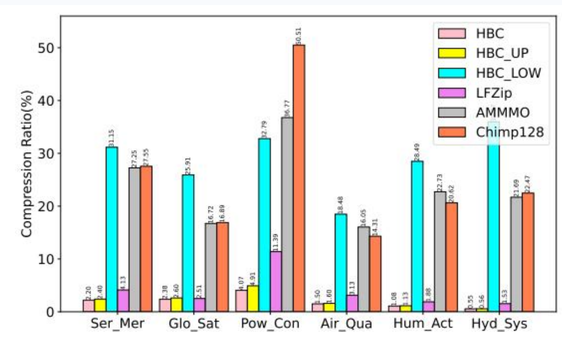


Fig. 5: Comparison of compression ratios for one-dimensional data.

TABLE V: Decompression time

	HBC	HBC_UP	HBC_LOW	LFZip	AMMMO	Chimp128
Ser_Mer	28.51	0.68	38.92	738.96	33.66	123.74
Glo_Sat	19.25	0.52	30.88	419.35	25.82	118.56
Pow_Con	30.89	1.26	38.56	870.12	34.49	130.42
Air_Qua	11.73	0.21	18.39	533.65	18.83	85.07
Hum_Act	14.66	0.39	20.24	613.77	19.27	98.96
Hyd_Sys	18.39	0.46	27.43	592.58	25.58	94.21
Average	20.57	0.59	29.07	628.07	26.28	108.49

TABLE IV: Compression time

	HBC	HBC_UP	HBC_LOW	LFZip	AMMMO	Chimp128
Ser_Mer	13.39	0.29	28.52	725.41	16.27	145.68
Glo_Sat	11.76	0.10	23.21	465.94	14.68	153.20
Pow_Con	19.85	0.56	30.74	943.03	23.22	139.14
Air_Qua	13.46	0.08	19.05	618.97	12.02	107.31
Hum_Act	14.53	0.14	22.53	560.47	13.85	112.65
Hyd_Sys	9.00	0.17	17.96	649.36	15.88	127.44
Average	13.66	0.23	23.67	660.53	15.99	130.90

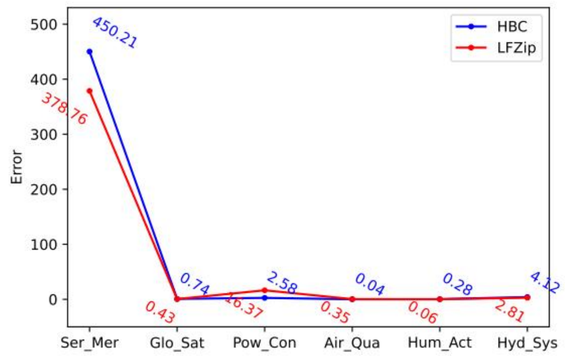


Fig. 7: Comparison of compression errors between HBC and LFZip.

ACKNOWLEDGMENT

This work is supported by the National Key R&D Program of China under No. 2021YFB2700500 and 2021YFB2700502, the Open Fund of Key Laboratory of Civil Aviation Smart Airport Theory and System, Civil Aviation University of China under No. SATS202206, the Open Fund of Key Laboratory of Complex Electronic System Simulation under No.614201002022205, the National Natural Science Foundation of China under No. U20B2050 and 82004499.