ADAPTIVE SENSOR DATA COMPRESSION IN IOT SYSTEMS: SENSOR DATA ANALYTICS BASED APPROACH

Arijit Ukil *Soma Bandyopadhyay* *Aniruddha Sinha* *Arpan Pal*

Innovation Labs
Tata Consultancy service, Kolkata, India
{arijit.ukil, soma.bandyopadhyay, aniruddha.s, arpan.pal} @tcs.com

ABSTRACT

Sensor nodes are embodiment of IoT systems in microscopic level. As the volume of sensor data increases exponentially, data compression is essential for storage, transmission and in-network processing. The compression performance to realize significant gain in processing high volume sensor data cannot be attained by conventional lossy compression methods. In this paper, we propose ASDC (Adaptive Sensor Data Compression), an adaptive compression scheme that caters various sensor applications and achieve high performance gain. Our approach is to exhaustively analyze the sensor data and adapt the parameters of compression scheme to maximize compression gain while optimizing information loss. We apply robust statistics and information theoretic techniques to establish the adaptivity criteria. We experiment with large sets of heterogeneous sensor datasets to prove the efficacy. Nonlinear lossy compression (Chebyshev) is extensively considered as the standard technique as well as experimental result with frequency domain compression like Discrete Fourier Transform (DFT) is shown as future scope of further improvement.

***Index Terms*—** Adaptive compression, information theory, IoT, sensor, Chebyshev

1. INTRODUCTION

With the proliferation of IoT (Internet of Things), sensors are omnipresent; from inside human body to deep inside Pacific Ocean. Such huge volumes of sensors generate high amount of data. Storage, transmission and processing of such high volume data pose potentially destruction performance and scalability risk. In order to harvest IoT applications like smart energy management, elderly monitoring, e-health care, data volume is to be reduced such that utility is optimized while maximizing the compression. Extensive research works already proposed various data compression methods particularly for sensor data [1, 2, 3]. Specifically, model-based approaches that compress sensor data by established approximation methods outperform traditional data compression methods [3]. Among various model-based compression techniques, the nonlinear model of Chebyshev approximation capture the inherent properties of large number of sensor data more accurately [4]. However, traditional notion of compression to minimize redundant information is not optimal for sensor data, as the contextual information also plays major role in analytics purpose. For example, unusual pattern in ECG data (may be indicative of arrhythmia) contain more analytic insight than regular ECG data; somebody's high energy consumption at odd hours (midnight) detected through smart meter readings is more interesting for mining and knowledge discovery. In continuation with that philosophy that is also supported by classical information theory, our proposed method ASDC extracts the useful or interesting information from sensor data and adapts the parameters like threshold selection, block-size estimation for Chebyshev compression to yield maximum compression gain while sacrificing insignificant information loss [17].

In this paper, we mainly focus on sensor data compression through Chebyshev polynomial. However, simpler techniques like frequency-domain compression using DFT (Discrete Frequency Transform) is also briefly studied, which would be our scope of future work.

The paper is organized as follows. In Section 2, we describe the sensor information extraction scheme. We briefly introduce Chebyshev approximation and describe our proposed scheme ASDC in Section 3. In Section 4, we discuss extension of ASDC to make it suitable for quasi-periodic signals like ECG. Results are shown in Section 5. Finally, we conclude in Section 6.

2. SENSOR DATA 'USEFUL' INFORMATION EXTRACTION

We follow the hypothesis that unusual pattern or anomalous events incur more useful information particularly in sensor data analytics, we employ robust outlier detection method to detect the interesting points in sensor datasets. Then, we quantify the amount of useful information as *sensor importance score*.

2.1. Useful Information Extraction

For robust outlier detection, masking effect and swamping effect need to be minimized [5, 8]. Available literature mostly concentrates on minimizing one of the effects [5, 7] due to the fact that a particular dataset (say stock market or smart meter reading) is prone to only one of the effects [19]. However, as we deal with heterogeneous (sensor) datasets, we need to minimize both the effects according to the dataset statistical properties [6]. Through kurtosis measurement (κ), we are able to estimate the spread of probability distribution $\mathcal{D}$ of sensor dataset $\varepsilon(t)$, which helps us to apply appropriate outlier detection technique**.**

2.2. Important Point Detection and Minimizing false alarms

In order to ensure minimum false alarm, both masking and swamping effects are to be minimized, It is shown in [8], Hampel identifier provides outstanding outlier detection when $\mathcal{D}$ can be approximated to random observation, which we interpret as $\kappa(\varepsilon) < 3$. Particularly MAD (Median Absolute Deviation)-analysis provides high neutralization to masking effects [8]. For minimizing swamping effect, which is predominantly observed when $\kappa(\varepsilon) \geq 3$, Rosner filtering is a judicious choice. Rosner filtering is a class of generalized extreme Studentized deviate (ESD) test for univariate multiple outlier detection [9]. It overcomes the limitation of *k*-means clustering, the Grubbs test and the Tietjen-Moore test by only requiring an upper bound rather a specific outlier suspected number *k* [9]. Consider ν be the useful information in sensor dataset ε and λ be the normal (uninteresting) part; $\varepsilon = \nu \cup \lambda$ and using Hampel/ Rosner filtering based on kurtosis measurement, we find out ν.

2.3. Quantification: Sensor Importance Score

Our idea of useful information measurement is the amount of difficulty to infer ν, when only given $\mathcal{S}$, i.e. the information leakage transfer function $\Upsilon_{\mathcal{S},\nu}$: $\mathcal{S} \rightarrow \nu$ and $\phi_M = \frac{\sum_{i=1}^{|\nu|} Pr(\nu_i) \log_2 \frac{1}{Pr(\nu_i)}}{\sum_{i=1}^{|S|} Pr(S_i) \log_2 \frac{1}{Pr(S_i)}} = \frac{H(\nu)}{H(S)}$ [10].

In [11], such metric is derived using mutual information $I(\mathcal{S},\nu)$, which would mostly indicate the maxima. In order to enhance the measurement accuracy, two-sample Kolmogorov-Smirnov (KS) test is performed. When KS-test accepts null hypotheses, statistical compensation $\rho_S = 1$. When KS-test rejects null hypotheses, we propose L1-Wasserstein metric between ε, ν $\left(\mathcal{w}_{\varepsilon,\nu}\right)$ to estimate statistical misfit or compensation $\phi_S = \mathcal{w}_{\varepsilon,\nu}$, where

$$\mathcal{w}_{\mathcal{S},\nu} \coloneqq \underset{\mu \in \Omega(\mathcal{S},\nu)}{\mathrm{INF}} \int_{\Omega} |x-y| \, d\mu(x,y),\ x \in \mathcal{S},\ y \in \nu$$

Logically, sensor importance score $(\phi_\varepsilon) = \phi_M \wedge \phi_S$. Algebraically, $\phi_\varepsilon = \phi_M \times \phi_S$. With ϕ_ε $[0, 5]$, we scale ϕ_ε as $\phi_\varepsilon \longmapsto [\phi_\varepsilon \times 5]$: $\phi_\varepsilon = [1, 5]$, with high magnitude of ϕ_ε signifies more useful information in ε.

3. FUNCTIONAL ARCHITECTURE OF ASDC

Below in fig. 1, we depict the functional architecture of ASDC, where adaptive module is the key for performance optimization and compression function $\mathcal{G} = f(\mathcal{P})$, where $\mathcal{P}$ is the set of adjustable parameters. For Chebyshev compression $\mathcal{G}_c = f(B,\ \Gamma)$, where B denotes the block size and Γ is the threshold value. Optimality in compression gain and information loss is achieved by adapting these two parameters [17]. Adaptation module as shown below consists of: 1. Block size optimization and 2. Threshold adaptation to deliver the optimized compression performance. For other types of compression technique like DFT, $\mathcal{P}$ consists of different elements like number of frequency components to be allowed.

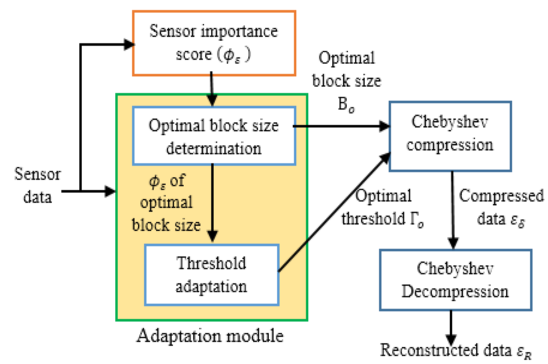

Figure 1. Functional architecture of Chebyshev-ASDC

3.1. Chebyshev Compression: A Brief

In Chebyshev compression, compressed data ε_δ is represented as linear combination of Chebyshev polynomials [3, 4]: $\varphi(i) = \sum_{i=0}^{i=N} \alpha_i . \eta_\vartheta(i)$, where: $\vartheta = \left(i - \frac{N+1}{2}\right) . \frac{2}{N-1}$, normalized to [-1, 1], simply, cosine look-up table, α_i is the Chebyshev polynomial co-efficient at degree *i*. Polynomial generation is done for every block B and for a defined threshold Γ [13]: $\varepsilon_\delta = \begin{cases} \varphi(i), & \varphi(i) \geq \Gamma \\ 0, & else \end{cases}$ (1) for each block. Quantization (digitization) is done for non-zero ε_δ for storage, transmission purpose.

3.2. Optimal Block Size

Conventional compression methods are designed upon the approximations of data under the norm L_∞.This maximum error constraint is violated in Chebyshev compression due to fixed pre-assigned block size B. We propose following algorithm to find the optimal block size B_o, which is dynamically adapted based on useful information differential $(\Delta\phi_\varepsilon)$ as:

$$B_o = \begin{cases} recursion\ (B_{init} = 2. B_{init}), & \Delta\phi_\varepsilon \neq 0 \\ B_{init} & else \end{cases}$$

Where, B_{init} is the initial block size and $\Delta\phi_{\varepsilon} = |\phi_{|B|} - \phi_{|B|+1}|$. The inherent effect of larger block size is to enhance compression gain as well as information loss. Whenever, differential useful information is non-zero for different block sizes, we consider larger block size. Here, we exploit the Asymptotic Equipartition Property (AEP) considering large number of samples available with (almost) equal probability of taking all the useful information score values at each block size and it helps us to achieve typical set.

3.3. Threshold Adaptation: Compression Gain-Information Loss Trade-off

Threshold (Γ) plays a major role in Chebyshev compression particularly for compression gain enhancement. Threshold acts as a clipping parameter as described in (1). Optimal determination of threshold is required for compression gain-information loss trade-off.

Let compression code $\varphi: \Theta \rightarrow \cup_{\pi \geq 1} \{0,1\}^{\pi}$ assigns each data points $\varepsilon\,(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N) \in \Theta\ (\Theta = \{0,1\}^{n})$, a finite sequence of *1*s and *0*s to create the (Chebyshev) code word $\varphi(\varepsilon)$. The condition for decompressibility of φ: $\sum_{\varepsilon \in \Theta} \mathrm{P}(\varepsilon)|\varphi(\varepsilon)| \geq -\sum_{\varepsilon \in \Theta} \mathrm{P}(\varepsilon) log_2 \mathrm{P}(\varepsilon)$, where, $\mathrm{P}(\varepsilon)$ is the probability distribution on the compression space Θ. From theory of large deviation, the optimality condition is : $\sum_{\varepsilon \in \Theta} \mathrm{P}(\varepsilon)|\varphi(\varepsilon)| = -\sum_{\varepsilon \in \Theta} \mathrm{P}(\varepsilon) log_2 \mathrm{P}(\varepsilon)$, when $\mathrm{P}(\varepsilon) = 2^{-NH(\varepsilon)} \approx 2^{-|\nu|H(\nu)}$ (according to the notion of useful information), which indicates that in order to achieve optimality in compression gain, Γ (Threshold) $\rightarrow \infty$. Large magnitude of Γ results the decompressed sequence would have type outside the set ε, i.e. the probability of information loss $(P_e) \rightarrow 1$. So the trade-off criteria is computed as: $\lim_{n \to \infty} \frac{1}{n} log_2 P_e(\varepsilon) = -KL(\varepsilon^D | \varepsilon)$ (from Sanov's theorem), where ε^D is the reconstruction (decoded) of raw signal ε and KL is the Kullback-Leibler divergence. Let's consider threshold (Γ): $\Gamma = c.2^{(\phi_{MAX} - \phi_{\varepsilon \in B_o})}$; As $\sum P_e(\varepsilon) = \sum_{n=1}^{N} \frac{1}{n} |\varepsilon^D - \varepsilon|$ and logarithm is a convex function, so, $\Gamma = c.2^{(\phi_{MAX} - \phi_{\varepsilon \in B_o})}$ is the optimal choice and constant factor c is dependent on dispersion factor to satisfy $P_e \rightarrow 0$. We consider $c = ceil(\mathcal{F})$, where $\mathcal{F}$ (Fano factor) $= \frac{\sigma^2}{\bar{\varepsilon}}$.

3.4. Results

Firstly, we depict the performance of ASDC for fixed block size ($B_o = 512$) with adaptive threshold as shown in fig. 2. All the experiments are done with publicly available sensor datasets like REDD [14], BLUED [15], Physionet [16]. Smart meter data is taken from REDD/ BLUED and ECG datasets are taken from Physionet. We consider smart meter datasets for initial experimentation. Five independent household smart meter data is chosen for first set of experimentation. We observe that significant performance gain in terms of lesser information loss and higher compression gain is achieved when block-size optimization (dynamic block-size) is employed (fig. 3) [17]. ASDC-const (ASDC with constant block size) Chebyshev compression provides better performance than standard Chebyshev, while block-size optimal Chebyshev (ASDC-adaptive) outperforms the ASDC-const.

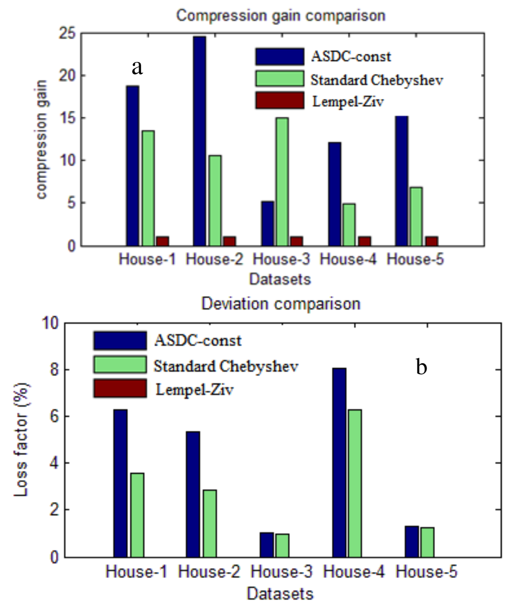

Figure 2. depicts (a) the compression gain, (b) loss factor variation of constant block-size Chebyshev-ASDC.

Compression gain $\left(\frac{\text{size of raw signal}}{\text{size of compressed signal}}\right)$ and loss factor % $\left(\frac{\text{std(compressed signal)} - \text{std(raw signal)}}{\text{std(raw signal)}}\right)$*100 for smart meter data, show considerable compression gain and lesser loss factor by adaptive ASDC.

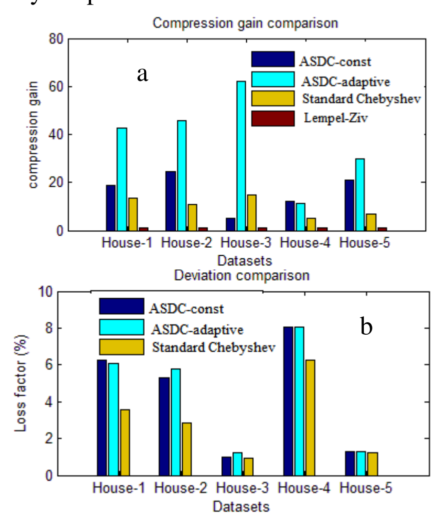

Figure 3. (a) Compression gain, (b) Loss factor, improvement with adaptive threshold along with optimal block size of Chebyshev-ASDC (ASDC-adaptive).

In fig. 4, original and reconstructed signal are shown.

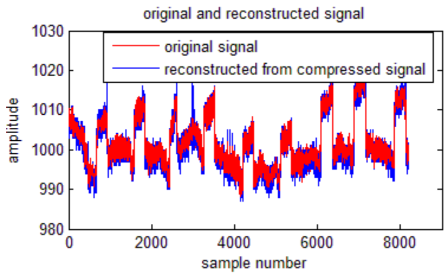

Figure 4. Original and reconstructed signal for House-1.

We extend our experiment with other different sensor signals as depicted in fig. 5, where EEG, ECG, accelerometer data are considered. We note that except ECG signal (signal indexed 6, 7, 8), others show good performance gain.

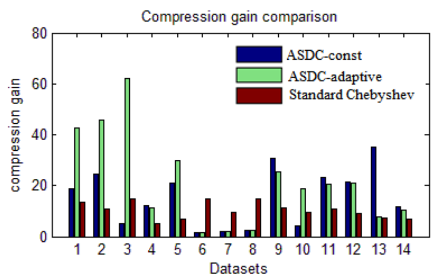

Figure 5. Performance gain comparison of ASDC for heterogeneous sensor data. It shows that few signals particularly of ECG class do not respond well.

4. EXTENTION OF ASDC

It is experimentally observed that sensor data with repeating sequences, i.e. signals with quasi-periodic components do not conform to the high performance gain of ASDC particularly due to the fact that the above mentioned derivation outlier is point-based. So the *useful information score* determination is not appropriate, more precisely outlier detection technique (Section 2.1) for quasi-periodic signals are prone to high false positive errors (swamping effect) when the hybrid approach of point-outlier detection is implemented. Consequently, we consider adaptive window-based discord discovery (AWDD) as the outlier detection technique [18] for quasi-periodic signals like ECG and rest of the analysis (Section 2.2 to 3.3) remains same. Below in fig. 6, we show the significant improvement of compression performance.

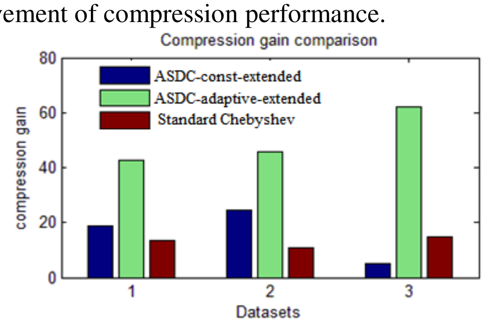

Figure 6. Improved compression performance with the help of extended ASDC with AWDD for quasi-periodic signals.

Truncation of Fourier series expansion (i.e. DFT in discrete domain) of the sensor data function is comprehensively produce compressed signal. By removing insignificant co-efficients ($|\text{co}-\text{efficients}|$ with $\leq \Gamma_{DFT}$), substantial compression gain can be achieved. However, the Γ_{DFT} (threshold) is to be dynamically adjusted to optimize the performance. Fig. 7 shows the compression gain and deviation loss for different class of sensor data types. We assumed constant Γ_{DFT} that makes compression gain similar but variation in loss. Our future scope of work is to derive the optimality on Γ_{DFT}. DFT-based compression would be simpler for practical implementation and suffices the real-time compression requirements for relevant applications.

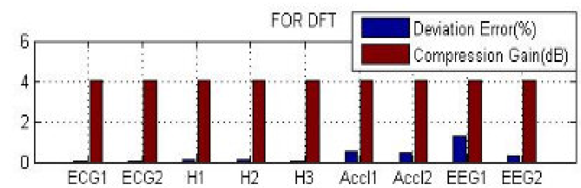

Figure 7. Compression performance of DFT-based compression.

5. CONCLUSION

In this paper, we proposed a novel sensor data compression scheme that outperforms conventional lossy compression scheme in significant margin. The proposed scheme Adaptive Sensor Data Compression (ASDC) exploits inherent statistical and information theoretic properties to maximize the compression gain while optimizing the data reconstruction loss. As, no single lossy compression provides better performance than others for every kinds of sensor datasets [3], our adaptive method of signal property based (if quasi-periodic, use discord discovery; else use robust point outlier technique) and information theoretic (Sanov's theorem based) approach proved to be the show stopper. We choose Chebyshev compression as the lossy compression technique, other lossy compressions like Grouping and Amplitude Scaling [1] can also be considered. We have already experimented with DFT-based compression, which will not only provide lower cost solution, but also its frequency-domain characteristics is independent of the intricacies of complex temporal analysis. Our future work is to experiment with larger set of sensor data to establish the universal efficacy of our proposed scheme with other relevant lossy compression techniques.

ACKNOWLEDGEMENT

We sincerely acknowledge Chetanya Puri, Researcher in Innovation Lab, Tata Consultancy Services, Kolkata for his contribution to the development and testing of AWDD algorithm.

6. REFERENCES

[1] S. Gandhi, et al. "GAMPS: compressing multi sensor data by grouping and amplitude scaling," SIGMOD, pp. 771- 784, 2009.

[2] X.T. Dang, N. Bulusu, and W. Feng, "RIDA: A robust information-driven data compression architecture for irregular wireless sensor networks," EWSN, pp. 133- 149, 2007.

[3] N. Hung, H. Jeung, and K. Aberer,"An evaluation of model-based approaches to sensor data compression," IEEE Transactions on Knowledge and Data Engineering, pp. 2434 - 2447, Nov. 2013.

[4] S. E. Hawkins III, and E. H. Darlington, "Algorithm for compressing time-series data," NASA Tech Briefs, pp. 7-8, June 2012.

[5] A. Ukil, S. Bandyopadhyay, and A. Pal, "Sensitivity Inspector: Detecting privacy in smart energy applications, " IEEE ISCC, 2014.

[6] R. Serfling, and S. Wang, "General foundations for studying masking and swamping robustness of outlier identifiers," Elsevier Statistical Methodology, August 2013.

[7] M. Gupta, J. Gao, C.C. Aggarwal, and H. Jiawei, "Outlier Detection for Temporal Data: A Survey," IEEE Transactions on Knowledge and Data Engineering, vol. 26, issue: 9, pp. 2250 - 2267, 2014.

[8] H. Liu, S. Shah, W. Jiang, "On-line outlier detection and data cleaning," Elsevier Computers and Chemical Engineering, pp. 1635–1647, 2004.

[9] B.Rosner, "Percentage points for a generalized ESD many-outlier procedure," Technometrics, vol. 25, issue. 2, pp. 165 - 172, 1983.

[10] A. Ukil, S. Bandyopadhyay, and A. Pal, "IoT privacy: To be private or not to be private," IEEE INFOCOM, 2014.

[11] L.Sankar, S.R. Rajagopalan, S. Mohajer, and H.V.Poor, "Smart meter privacy: A theoretical framework," IEEE Transactions on Smart Grid, vol. 4, issue. 2, pp. 837 - 846, 2013.

[12] F. Krzakala, M. Mezard, and L. Zdeborova, "Compressed sensing under matrix uncertainty: optimum thresholds and robust approximate message passing," IEEE ICASSP, pp. 5519 - 5523, 26 - 31 May, 2013.

[13] R. Banerjee, A. Sinha and K. Chakravarty, "Gait based people identification system using multiple switching kinects," ISDA, December 8 - 10, 2013.

[14] Z. Kolter, and M.J. Johnson, "REDD: A public data set for energy disaggregation research," ACM SustKDD, 2011.

[15] K. Anderson, et al. "BLUED: A fully labeled public dataset for event-based non-intrusive load monitoring research," ACM SustKDD, 2012.

[16] D. Guo, el al. "ECG-derived cardiopulmonary analysis of pediatric sleep-disordered breathing," Sleep Medicine, pp. 384-389, 2011.

[17] A. Ukil, S. Bandyopadhyay, and A. Pal, "IoT Data Compression: Sensor-Agnostic Approach," Data Compression Conference, pp. 303- 312, 2015.

[18] F.C. Chuah, and F. Fu, "ECG anomaly detection via time series analysis," ACM ISPA, pp. 123 – 135, 2007.

[19] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier Detection Techniques for Wireless Sensor Networks: A Survey," IEEE Communications Surveys & Tutorials, vol. 12, Issue: 2, pp. 159 - 170, 2010.