DASlab

FUDAN UNIVERSITY

Tsinghua University

Akane: Perplexity-Guided Time Series Data Cleaning

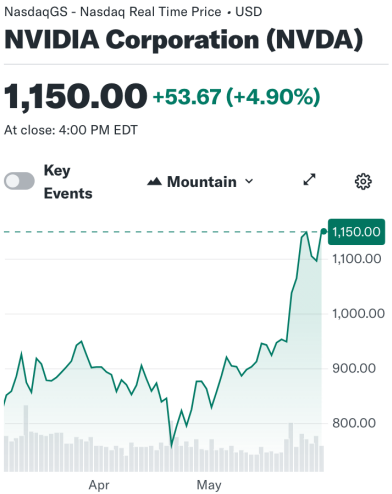
2024.06.11

Xiaoyu Han, Haoran Xiong, Zhenying He, Peng Wang, Chen Wang, X. Sean Wang

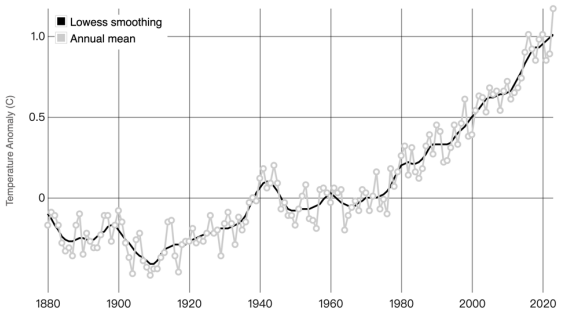
Background

DASlab
daslab.fudan.edu.cn

- Time series analysis
  - Finance, Meteorology, Industry, Medicine, etc.



Yahoo Finance



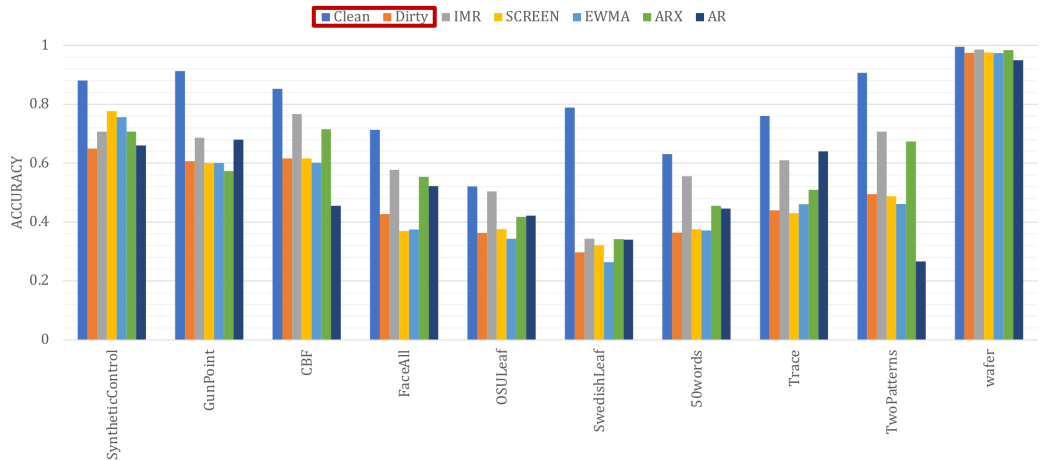
NASA Global Temperature

Background

DASlab
daslab.fudan.edu.cn

- Premise of reliable analysis
  - Good data quality

-- Data quality issues have been widely recognized in IoT data, and prevent the downstream applications. [CIKM'20]



Background

- Time series reliability issues
  - Unreliable devices
  - Malfunctioning metering systems
  - Confuse of stock symbols

Background

- Simply discarding might not work
  - Sometimes data are precious
  - Make data even more incomplete [CIKM'20]
  - Mislead downstream applications [CIKM'20]

Background

- Simply discarding might not work
  - Sometimes data are precious
  - Make data even more incomplete [CIKM'20]
  - Mislead downstream applications [CIKM'20]

Time series data cleaning

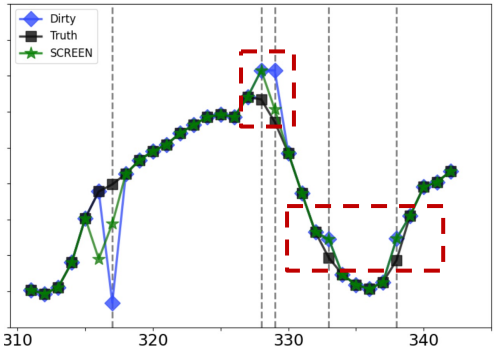
Motivation

- Classical algorithms
  - Moving Average (MA)
    - Change too many points
  - Auto Regression (AR)
    - Fit imprecisely

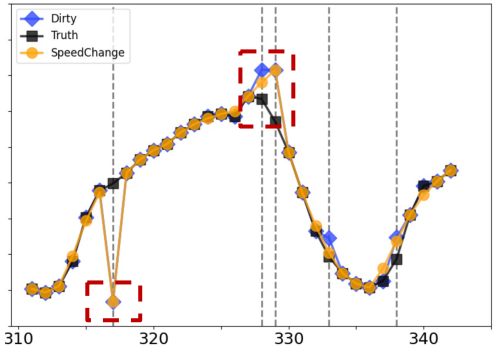
Motivation

- Prevalent algorithms
  - SCREEN [SIGMOD'15]
  - SpeedChange [SIGMOD'16]
  - ~~IMR [VLDB'17] (supervised cleaning)~~
  - Numerous anomaly-detection based methods
  - …

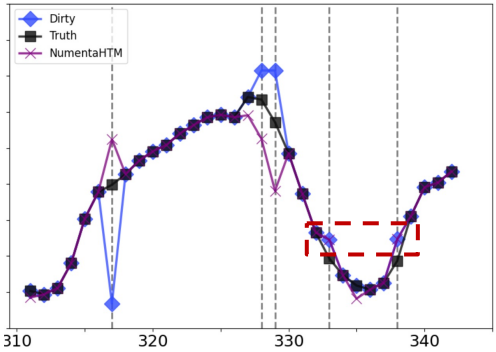
Motivation



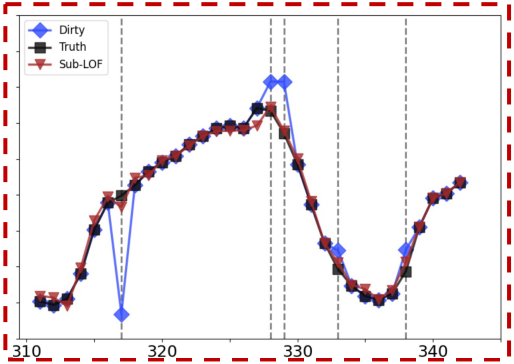
SCREEN



SpeedChange



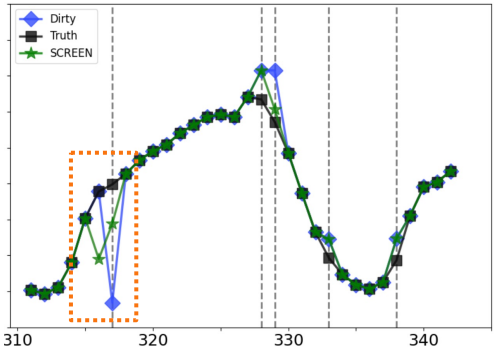
NumentaHTM



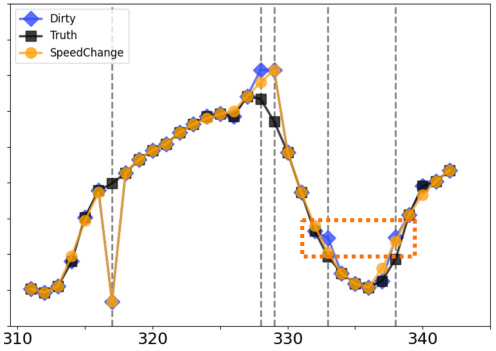
Sub-LOF

Mistakes and omissions in dirty data identification

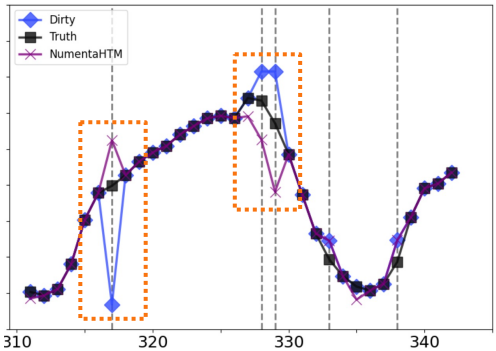
Motivation



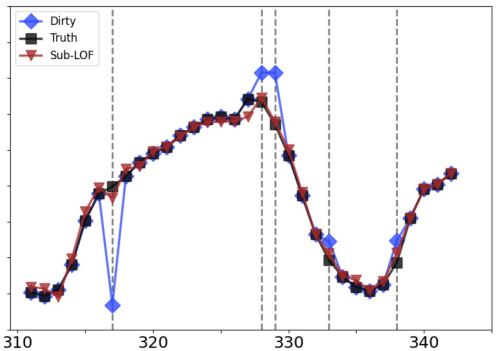
SCREEN



SpeedChange



NumentaHTM

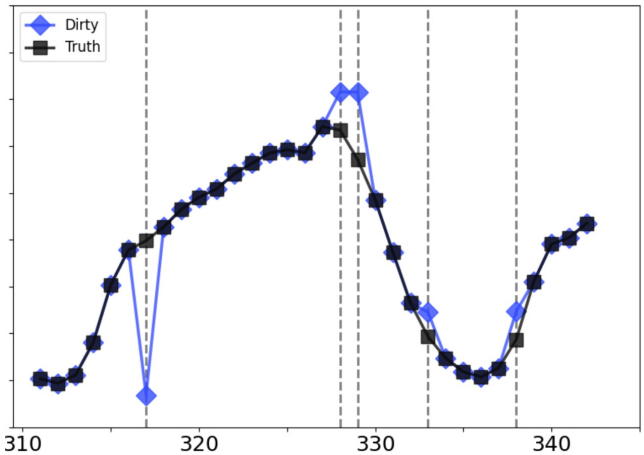


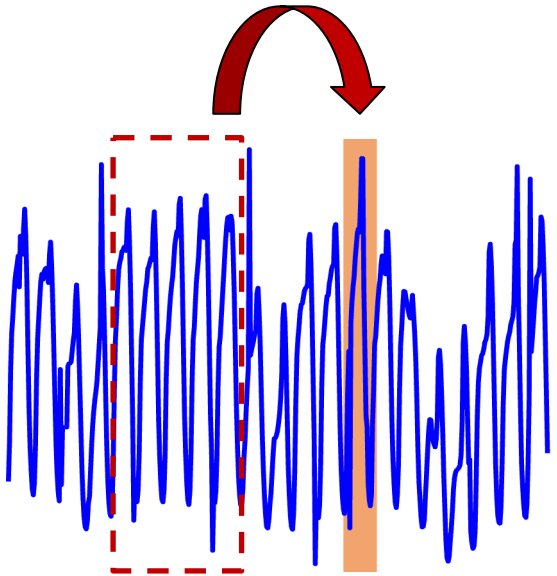
Sub-LOF

Deviate a lot from the truth in cleaning decisions

Motivation

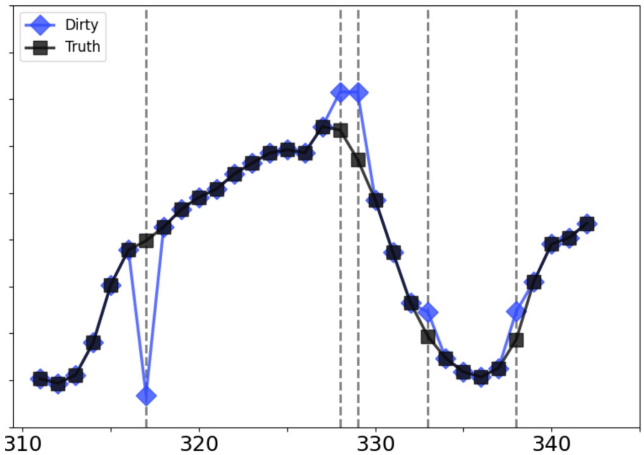
- Why fail?

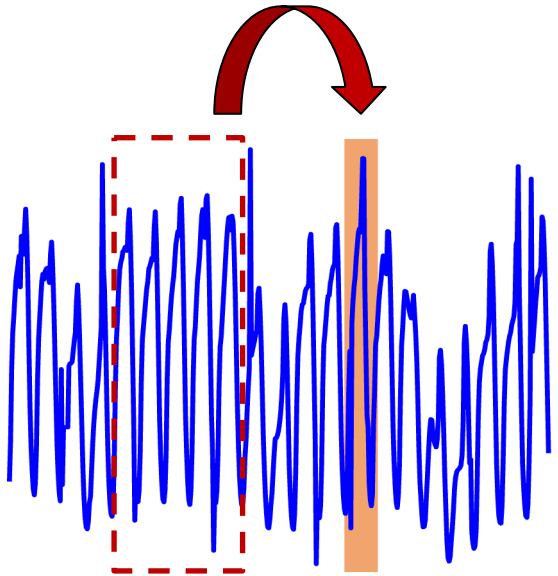




Motivation

- Why fail?

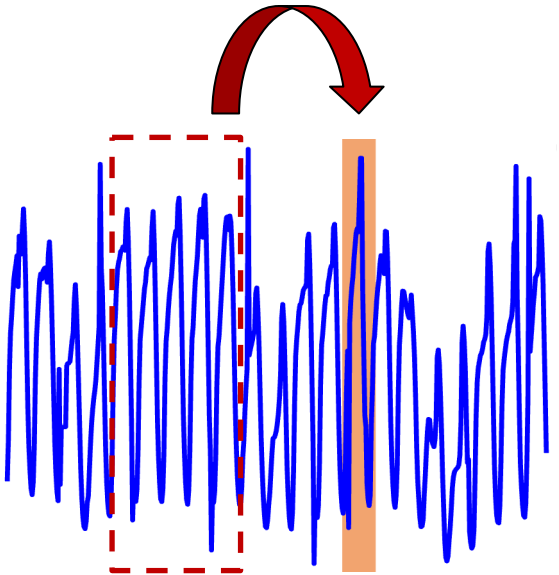




Plight: **Limited use** of inherent time series information

Intuition

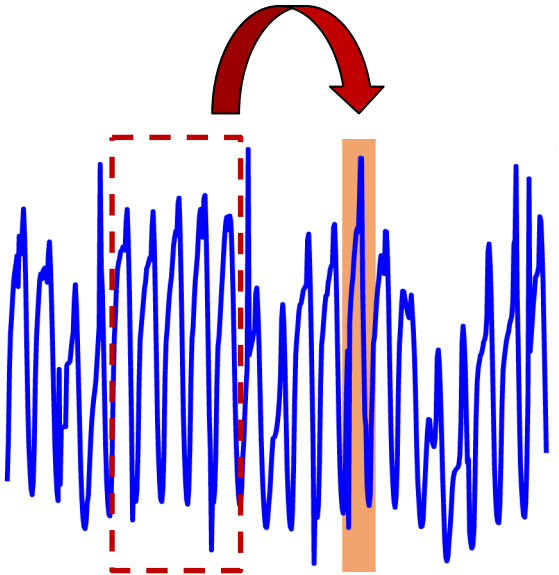
- What information can we use?



What are these?

Intuition

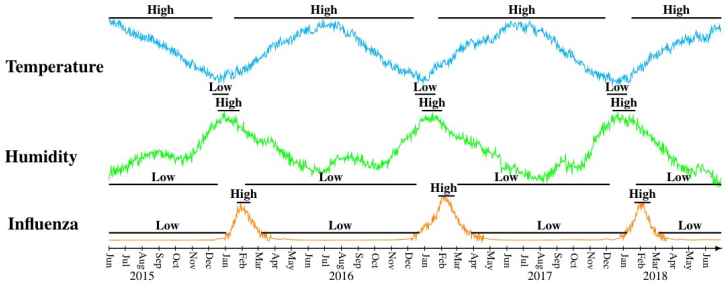
- What information can we use?



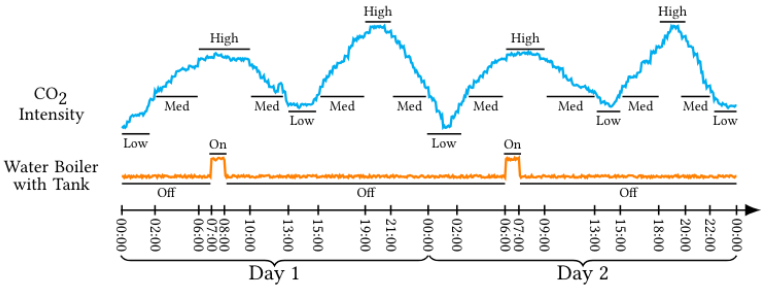
Recurrent patterns!

Intuition

- Recurrent patterns in the time series



Weather and Influenza time series [ICDE'23]



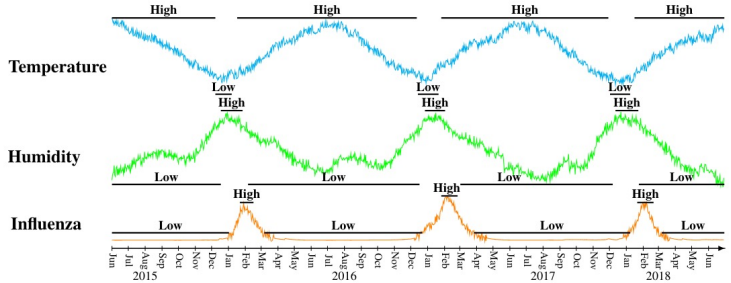
CO2 intensity and water boiler electricity usage [VLDB'21]

Intuition

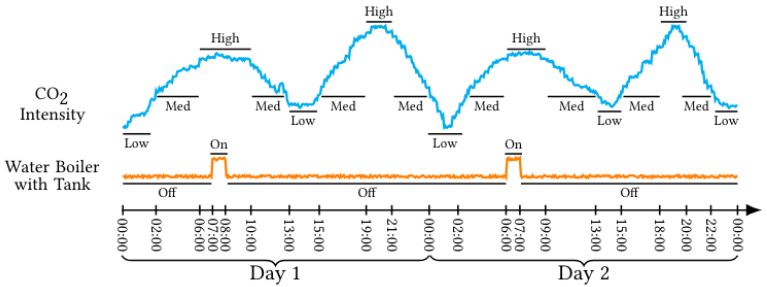
DASlab
daslab.fudan.edu.cn

- Recurrent patterns in the time series

Recurrent ≠ Seasonal



Weather and Influenza time series [ICDE'23]



CO2 intensity and water boiler electricity usage [VLDB'21]

Intuition

- How to use recurrent patterns?
- How to use similar patterns elsewhere in the time series to clean a point?

Intuition

- Analogize recurrent patterns in a time series to fixed word combinations in a sentence

Intuition

- Analogize recurrent patterns in a time series to fixed word combinations in a sentence
  - Repeated occurrence
  - Sequential dependency
  - A small part of whole data

Intuition

- Perplexity in NLP
  - An evaluation metric that measures the quality of language models, especially for text generation

$$PP(W) = P(w_1, w_2, \cdots, w_N)^{-\frac{1}{N}}$$

Intuition

- Perplexity in NLP
  - An evaluation metric that measures the quality of language models, especially for text generation

$$PP(W) = P(w_1, w_2, \cdots, w_N)^{-\frac{1}{N}}$$

Introduce it to evaluate time series data quality

Lower perplexity, better quality

Problem Definition

- Perplexity-Guided time series data cleaning
  - Given a dirty time series, cleaning the time series equals to reducing its perplexity within the budget

Problem Definition

- Perplexity-Guided time series data cleaning
  - Given a dirty time series, cleaning the time series equals to reducing its perplexity within the budget

Cost: $\Delta(X, X') = \| \{x_i \,|x_i \neq x_i', 1 \leq i \leq n\} \|$

Problem Definition

DASlab
daslab.fudan.edu.cn

- Perplexity-Guided time series data cleaning
  - Given a dirty time series, cleaning the time series equals to reducing its perplexity within the budget

?

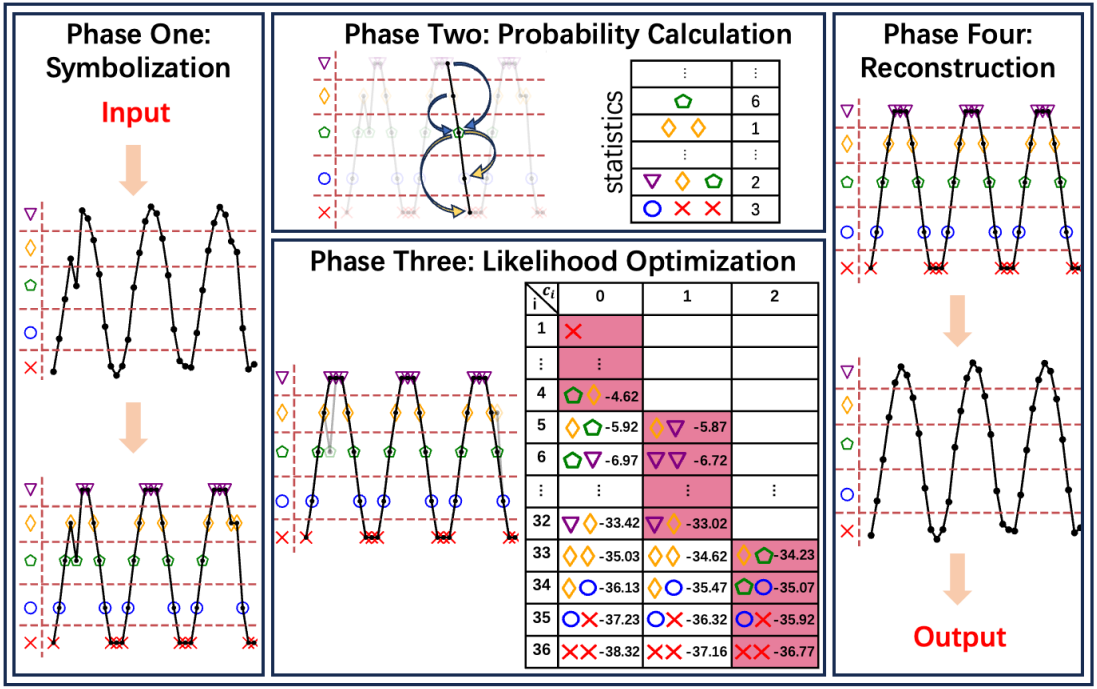
continuous value

Cost: $\Delta(X, X') = \| \{x_i \,|x_i \neq x_i', 1 \leq i \leq n\} \|$

Framework

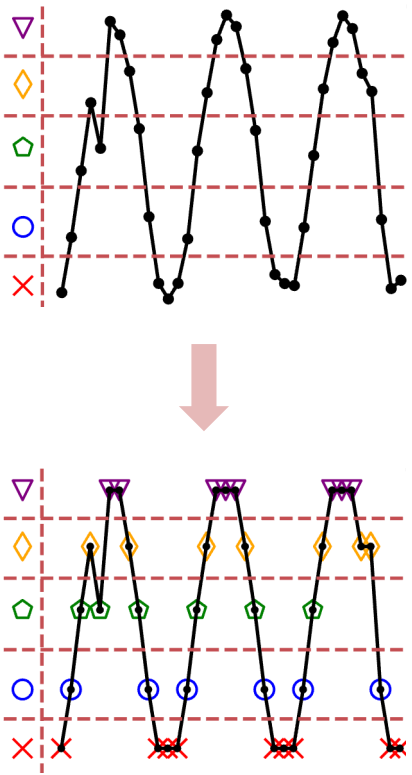
DASlab
daslab.fudan.edu.cn

- Overview



Framework

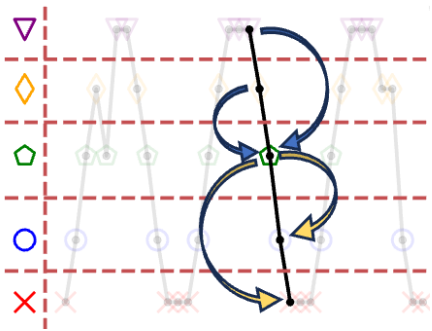
- Symbolization
  - Change continuous value to discrete symbols to calculate the perplexity like textual data
  - K-means / uniform symbolization

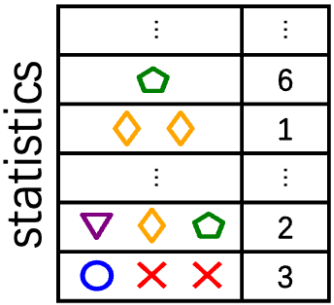


Framework

- Probability calculation
  - Use statistics (combinations) to calculate occurrence probability
  - Introduce *k*-order Markov chain

$$P(x_i^w | x_{i-k:i-1}^w) = \left(C\left(x_{i-k:i}^w\right)\right) / \left(C\left(x_{i-k:i-1}^w\right)\right)$$

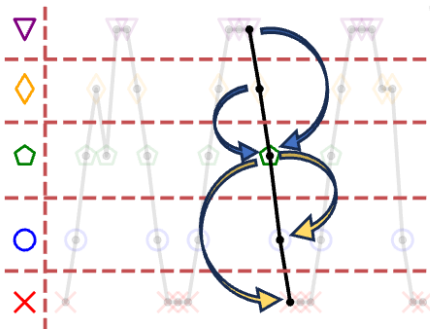


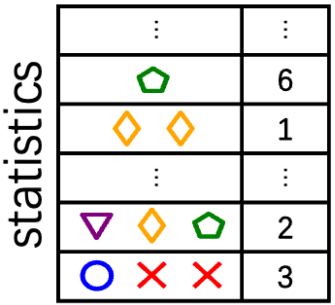


Framework

- Probability calculation
  - Use statistics (combinations) to calculate occurrence probability
  - Introduce *k*-order Markov chain

$$P(x_i^w | x_{i-k:i-1}^w) = \left(C\left(x_{i-k:i}^w\right) + 1\right) / \left(C\left(x_{i-k:i-1}^w\right) + r\right)$$





Framework

- Likelihood optimization
  - Use likelihood rather than perplexity

$$PP(X) = PP(X^w) = \sqrt[n]{1/P(X^w)}$$

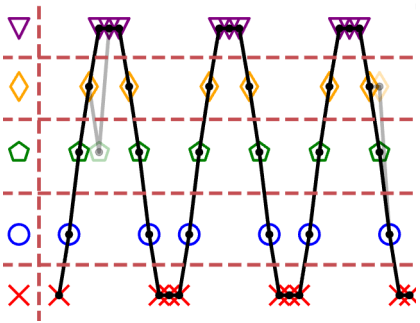
$$L(X) = L(X^w) = \log P(X^w)$$

Framework

- Likelihood optimization
  - Aim to maximize likelihood within budget

Framework

- Likelihood optimization
  - A knapsack-like problem
  - Use pseudo-polynomial time algorithm based on Dynamic Programming



i \ c_i	0	1	2
1	✕		
⋮	⋮		
4	-4.62		
5	-5.92	-5.87	
6	-6.97	-6.72	
⋮	⋮	⋮	⋮
32	-33.42	-33.02	
33	-35.03	-34.62	-34.23
34	-36.13	-35.47	-35.07
35	-37.23	-36.32	-35.92
36	-38.32	-37.16	-36.77

Framework

- Likelihood optimization
  - Recurrence equation

$$\Lambda\left(i, c_i, x'^{w}_{i-k+1:i}\right) = \max_{x'^{w}_{i-k} \in W_s} \left[\Lambda\left(i-1, c_{i-1}, x'^{w}_{i-k:i-1}\right) + \log P\left(x'^{w}_{i} | x'^{w}_{i-k:i-1}\right)\right]$$

  - $\Lambda(i, c_i, x'^{w}_{i-k+1:i})$ is the maximum likelihood of cleaned $x'^{w}_{1:i}$ whose last k points are $x'^{w}_{i-k+1}, \cdots, x'^{w}_{i}$
  - $c_i = \Delta(x'^{w}_{1:i}, x^{w}_{1:i})$ is the corresponding cleaning cost

Framework

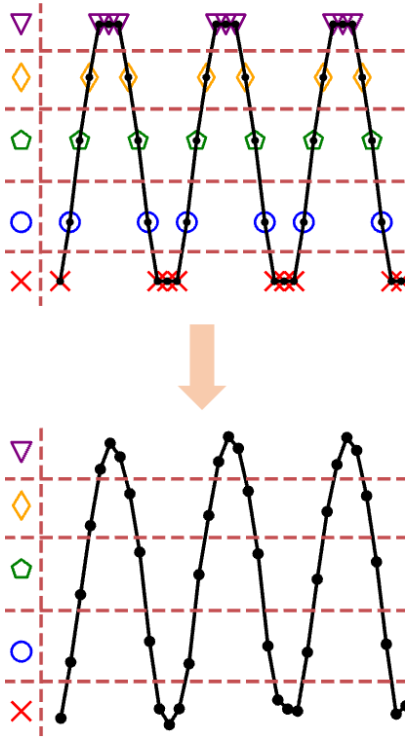
- Likelihood optimization
  - Recurrence equation

$$\Lambda\left(i, c_i, x'^{w}_{i-k+1:i}\right) = \max_{x'^{w}_{i-k} \in W_s} \left[\Lambda\left(i-1, c_{i-1}, x'^{w}_{i-k:i-1}\right) + \log P\left(x'^{w}_{i} | x'^{w}_{i-k:i-1}\right)\right]$$

  - Get cleaned result through retracement
  - Time complexity $O\left(r^{k+1}kn^2\right), r = |W_s|$
  - Space complexity $O(r^k n^2)$

Framework

- Reconstruction
  - Change discrete symbols back to continuous values
  - For each cleaned point, fit Linear Regression (LR) to obtain the most possible value

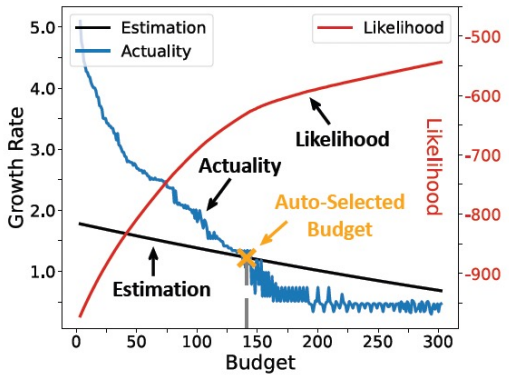


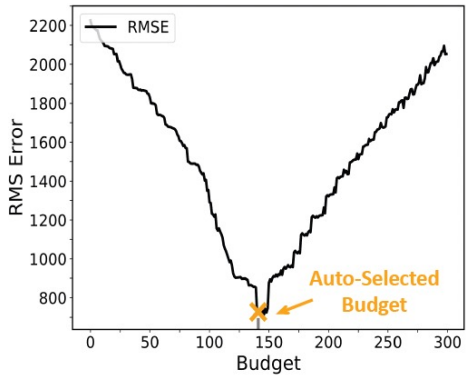
Framework

- Parameter advice
  - K-means symbolization number r
    - argmin Davies-Bouldin Index
  - Uniform symbolization number r
    - estimated range $[\frac{X_{max}-X_{min}}{2*|\bar{v}|}, \frac{2*(X_{max}-X_{min})}{|\bar{v}|}]$
  - Markov chain order k
    - directly choose 2 or 3

Framework

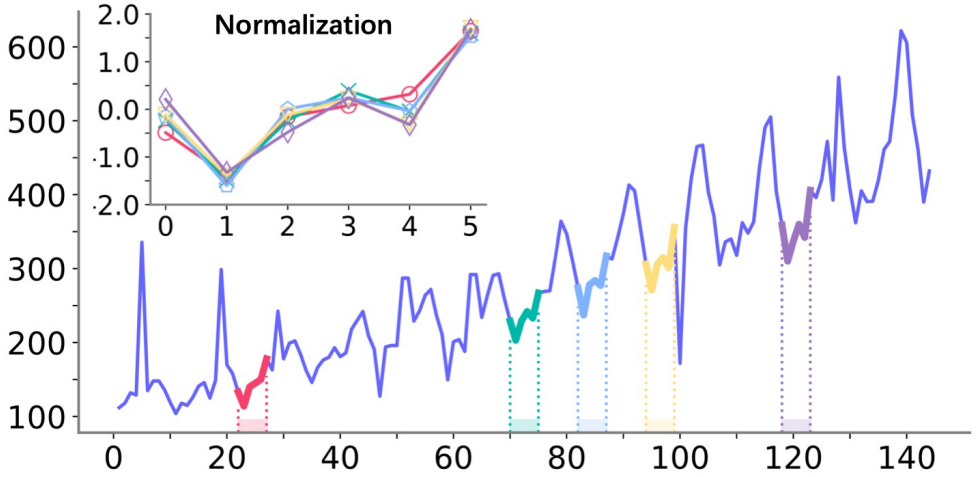
- Parameter advice
  - Cleaning budget δ
    - chase actual likelihood increase till $\log \frac{m * (1-\tau)^{k+1}+1}{m * (1-\tau)^{k} * \frac{\tau}{r-1}+1}$
    - τ is current estimated dirty rate (we clean gradually)





Advanced Solutions

- Homomorphic pattern
  - Different patterns but behaving similarly



Advanced Solutions

- Homomorphic pattern aggregation
  - Aggregate similar patterns to a homomorphic set

$$Dist\left(\hat{x}^{w}_{p-k:p}, \hat{x}^{w}_{q-k:q}\right) \leq \epsilon$$ (distance after normalization)

Advanced Solutions

- Homomorphic pattern aggregation
  - Aggregate similar patterns to a homomorphic set

$$Dist\left(\hat{x}^{w}_{p-k:p}, \hat{x}^{w}_{q-k:q}\right) \leq \epsilon$$ (distance after normalization)

An empirical value within 0.08 is good

Advanced Solutions

- Homomorphic pattern aggregation
  - Aggregate similar patterns to a homomorphic set

$$Dist\left(\hat{x}^{w}_{p-k:p}, \hat{x}^{w}_{q-k:q}\right) \leq \epsilon$$ (distance after normalization)

  - NP-Hard to find minimum sets from a time series
  - We propose a heuristic method to solve it

Advanced Solutions

- Homomorphic pattern aggregation
  - Use set frequency to calculate probability
  - For a pattern $x'^{w}_{p_i-k:p_i}$ in a homomorphic set $\{x'^{w}_{p_1-k:p_1}, \cdots, x'^{w}_{p_s-k:p_s}\}$ of cardinality s

$$P\left(x'^{w}_{p_i} \middle| x'^{w}_{p_i-k:p_i-1}\right) = \frac{\sum_{j=1}^{s} C\left(x'^{w}_{p_j-k:p_j}\right) + 1}{\sum_{j=1}^{s} C\left(x'^{w}_{p_j-k:p_j-1}\right) + r}$$

Advanced Solutions

- Homomorphic pattern aggregation
  - Use set frequency to calculate probability
  - For a pattern $x'^{w}_{p_i-k:p_i}$ in a homomorphic set $\{x'^{w}_{p_1-k:p_1}, \cdots, x'^{w}_{p_s-k:p_s}\}$ of cardinality s

$$P\left(x'^{w}_{p_i} \middle| x'^{w}_{p_i-k:p_i-1}\right) = \frac{\sum_{j=1}^{s} C\left(x'^{w}_{p_j-k:p_j}\right) + 1}{\sum_{j=1}^{s} C\left(x'^{w}_{p_j-k:p_j-1}\right) + r}$$

Advanced Solutions

- Homomorphic pattern aggregation
  - Use set frequency to calculate probability
  - For a pattern $x'^{w}_{p_i-k:p_i}$ in a homomorphic set $\{x'^{w}_{p_1-k:p_1}, \cdots, x'^{w}_{p_s-k:p_s}\}$ of cardinality s

$$P\left(x'^{w}_{p_i} \middle| x'^{w}_{p_i-k:p_i-1}\right) = \frac{\sum_{j=1}^{s} C\left(x'^{w}_{p_j-k:p_j}\right) + 1}{\sum_{j=1}^{s} C\left(x'^{w}_{p_j-k:p_j-1}\right) + r}$$

Time complexity $O\left(r^{k+1}kn^2\right)$ ⇨ $O(r^{k+1}kn^2 \log n)$

Advanced Solutions

- Greedy-based heuristic algorithm
  - Global optimum is too consuming
  - Each time we choose to clean one point to obtain maximum likelihood increase

$$\arg\max_{i,x_i''^w} \sum_{j=i}^{i+k} \left[\log P(x_j''^w | x''^w_{j-k:j-1}) - \log P(x_j'^w | x'^w_{j-k:j-1}) \right]$$

Advanced Solutions

- Greedy-based heuristic algorithm
  - Global optimum is too consuming
  - Each time we choose to clean one point to obtain maximum likelihood increase

$$\arg\max_{i,x_i''^w} \sum_{j=i}^{i+k} \left[\log P(x_j''^w | x''^w_{j-k:j-1}) - \log P(x_j'^w | x'^w_{j-k:j-1}) \right]$$

No aggregation $O\left(r^{k+1}kn^2\right)$ ⇨ $O(rk^3n + kn\log n)$

With aggregation $O(r^{k+1}kn^2\log n)$ ⇨ $O(rk^3n\log n)$

Experiments

- Datasets
  - 12 datasets with synthetic errors and real errors

Name	Dirty Type	Length	Rate	Source	Notes
CA	Synth	1K-10K	5-30%	California ISO	Energy
Romania	Synth	1K-10K	5-30%	Kaggle	Energy
Product	Synth	397	20%	FRED	Utilities
Retail	Synth	377	20%	FRED	Trade
Passenger	Synth	144	20%	Kaggle	Airline
Traffic	Real	200	2%	Dodgers [3, 26]	Flow
ID_7c18	Real	1500	0.73%	IOPS [36]	KPI
ID_7698	Real	1K	0.2%	IOPS [36]	KPI
ID_a40b	Real	10K	0.32%	IOPS [36]	KPI
MSL-T-5	Real	2218	0.23%	NASA-MSL [25]	Telemetry
SMAP-G-2	Real	7361	0.014%	NASA-SMAP [25]	Telemetry
Stock	Synth	12824	10%	SCREEN [41]	Finance

Experiments

- Algorithms
  - Akane, AkaneH, AkaneH+
  - EWMA, AR-Linear, AR-LSTM, SCREEN, SpeedChange, Torsk, NumentaHTM, TripleES, FFT, GrammarViz, Sub-LOF

Experiments

- Algorithms
  - Akane, AkaneH, AkaneH+ → K-means
  - EWMA, AR-Linear, AR-LSTM, SCREEN, SpeedChange, Torsk, NumentaHTM, TripleES, FFT, GrammarViz, Sub-LOF

Experiments

- Algorithms
  - Akane, AkaneH, AkaneH+
  - EWMA, AR-Linear, AR-LSTM, SCREEN, SpeedChange, Torsk, NumentaHTM, TripleES, FFT, GrammarViz, Sub-LOF

uniform + aggregation

Experiments

- Measurements
  - Root Mean Square Error (RMSE)
  - Accuracy

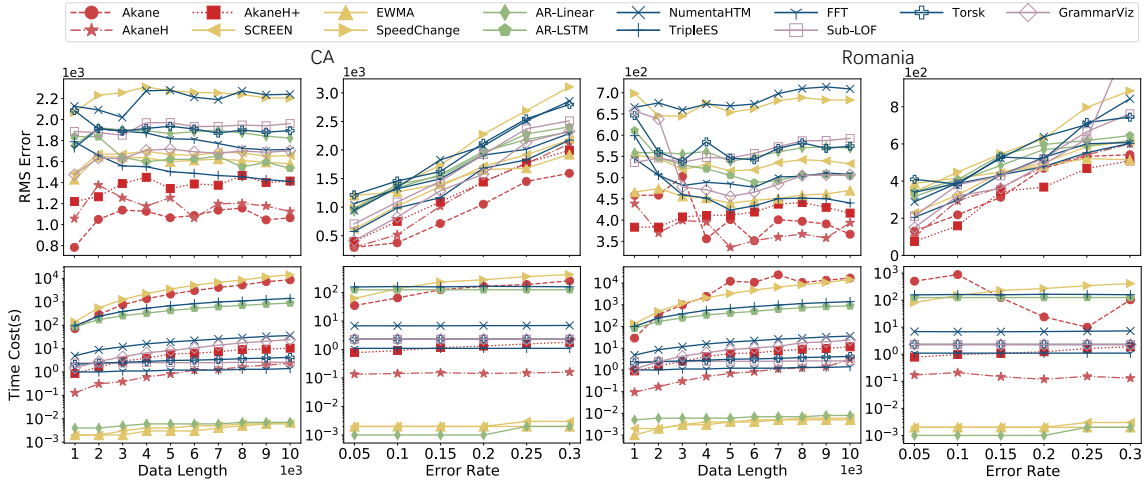
$$accuracy = \left(RMSE^{dirty} - RMSE^{cleaned}\right) \Big/ \left(RMSE^{dirty} - RMSE^{cleaned}_{min}\right)$$

  - Execution time cost

Experiments

DASlab
daslab.fudan.edu.cn

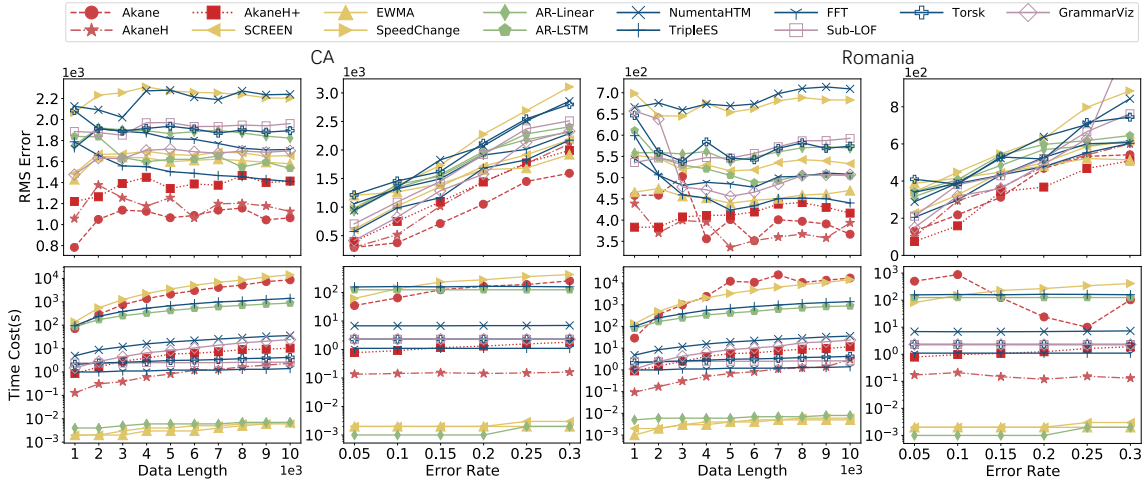
- Vary length & error rate on CA and Romania



Experiments

DASlab
daslab.fudan.edu.cn

- Vary length & error rate on CA and Romania

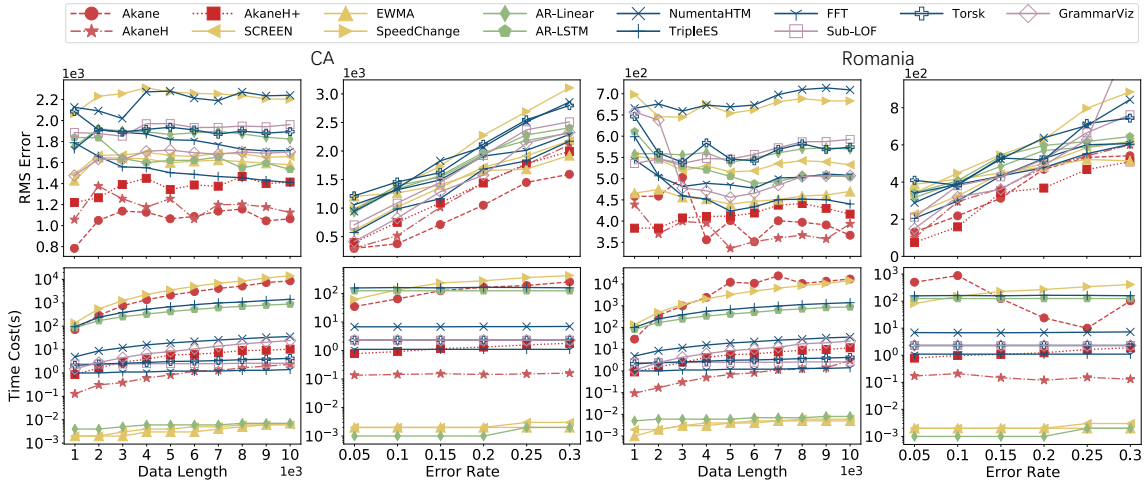


1. More powerful than existing algorithms

Experiments

DASlab
daslab.fudan.edu.cn

- Vary length & error rate on CA and Romania

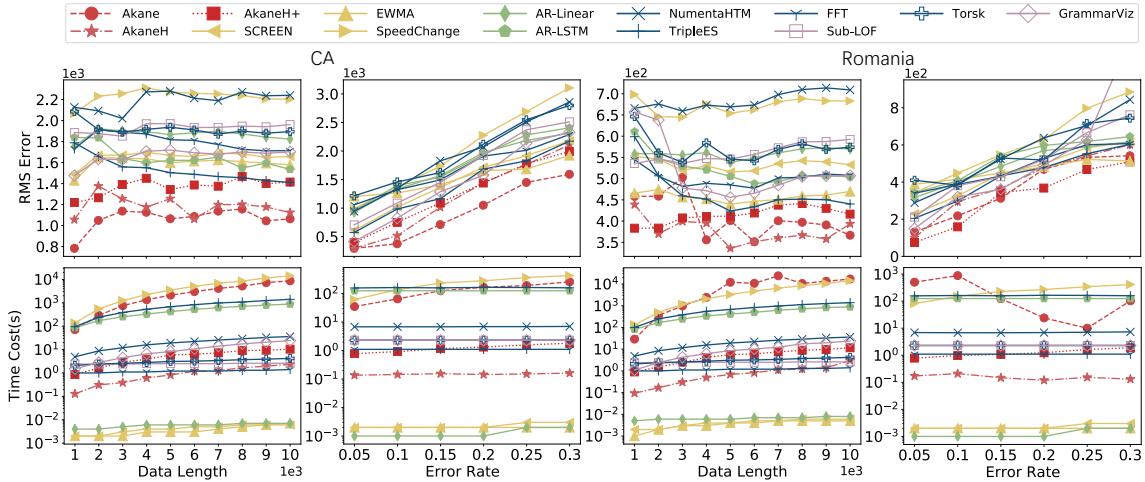


1. More powerful than existing algorithms
2. Relatively stable at all length; Better at low rate

Experiments

DASlab
daslab.fudan.edu.cn

- Vary length & error rate on CA and Romania

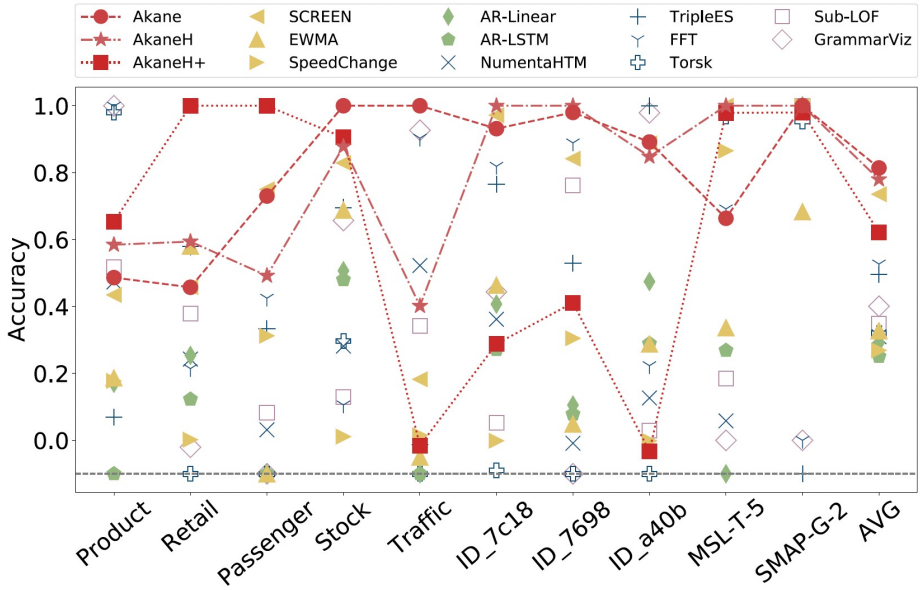


1. More powerful than existing algorithms
2. Relatively stable at all length; Better at low rate
3. Faster heuristic algorithm without many sacrifices

Experiments

DASlab
daslab.fudan.edu.cn

- Performance on various datasets

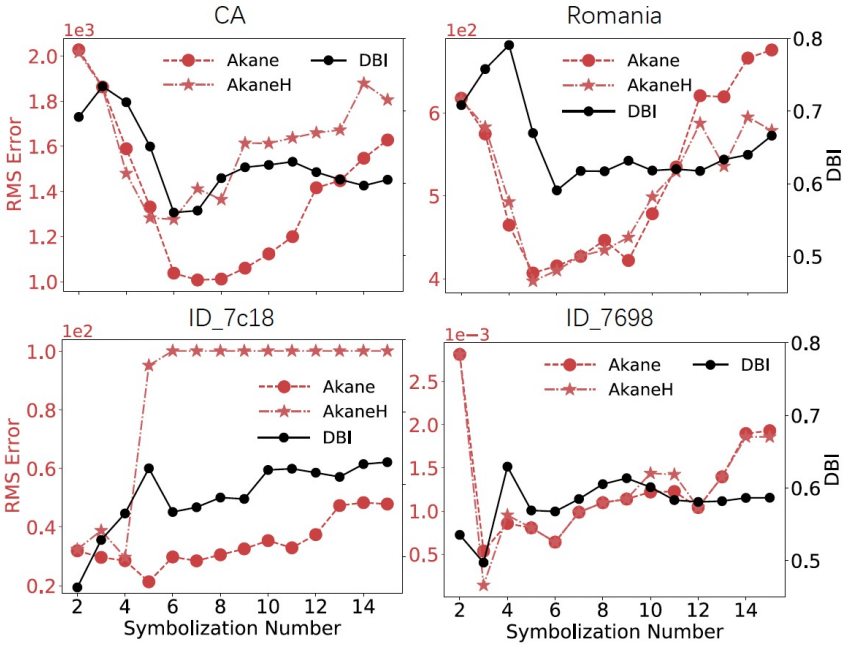


Overall great performance

Experiments

DASlab
daslab.fudan.edu.cn

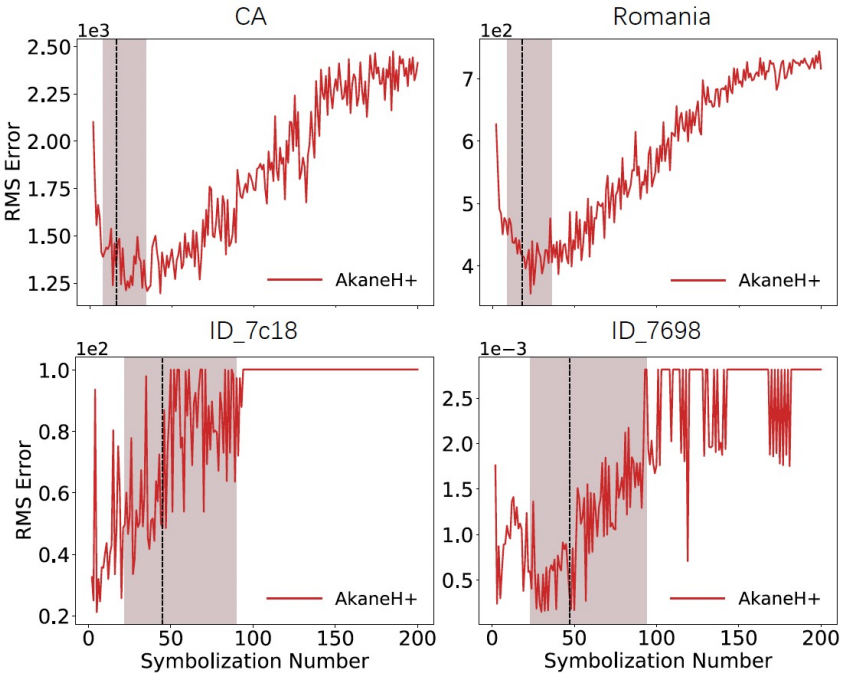
- Vary K-means symbolization number r



Experiments

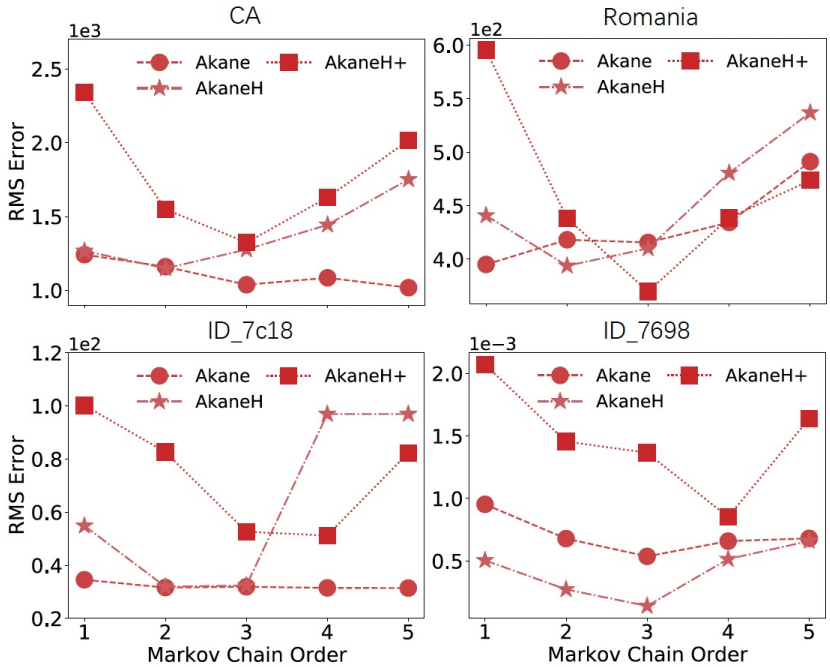
DASlab
daslab.fudan.edu.cn

- Vary uniform symbolization number r



Experiments

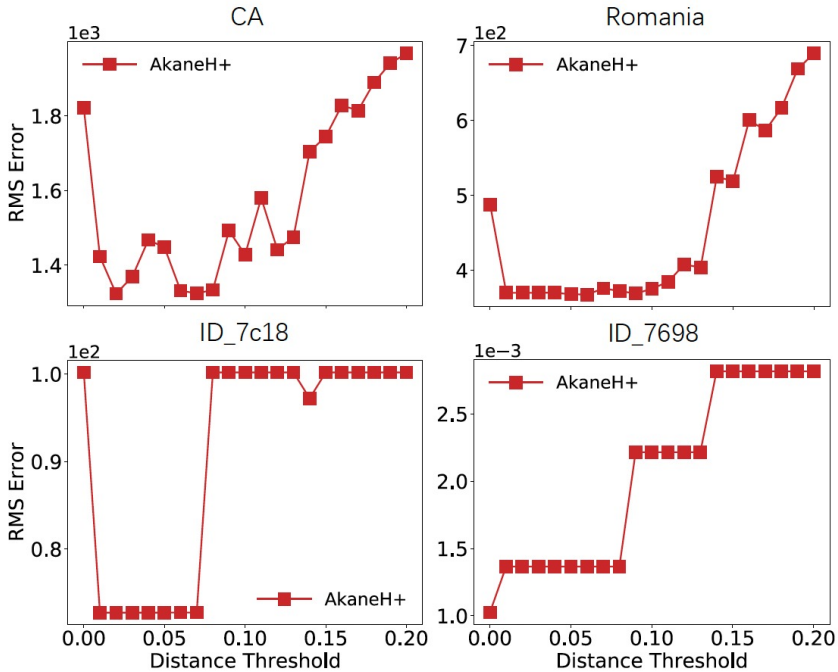
- Vary Markov chain order k



Experiments

DASlab
daslab.fudan.edu.cn

- Vary distance threshold ϵ



Experiments

- Auto budget selection strategy evaluation

Evaluation	Akane	AkaneH	AkaneH+
Vary CA Length	0.953	0.944	0.966
Vary Romania Length	0.922	0.947	0.973
Vary CA Rate	0.974	0.894	0.939
Vary Romania Rate	0.892	0.853	0.991
10 fixed datasets	0.829	0.833	0.811

Experiments

- Parameter experiments show
  - Performances have close relations to parameters
  - Our parameter advice is effective

Conclusions

- We use the insight of recurrent patterns to formalize perplexity-guided time series data cleaning problem
- We propose a four-phase framework to solve the cleaning problem, with parameter analysis and advice
- We further introduce advanced solutions, involving homomorphic pattern aggregation and greedy-based heuristic algorithm, to enhance generalization

THANKS!