National and Kapodistrian
University of Athens

A Spatio-Temporal Data Imputation Model for Supporting Analytics at the Edge

Kostas Kolomvatsos, **Panagiota Papadopoulou**, Christos Anagnostopoulos, Stathes Hadjiefthymiades

The 18th IFIP Conference on E-Business, E-Services and E-Society (I3E 2019)
Trondheim, Norway
18-20 September 2019

Outline

Edge Computing

Numerous Devices

The Internet of Things (IoT) infrastructure
End users carry their own device
Devices can be interconnected

Huge Volumes of Data

IoT devices collect data
Applications produce data
Data should be managed to provide analytics

Interaction with the Edge of the Network

Edge nodes are placed close to the data collection
They can serve analytics queries on top of the collected data
Edge computing minimizes the latency in the provision of responses

Data Imputation

Data streams can involve missing values
Such values should be managed
Various methodologies have been proposed

Our Contribution

State of the Art	Our focus	We propose
Legacy techniques mainly focus on the statistics of data	We incorporate the dynamics of the environment where IoT devices act	We combine the reports of multiple IoT devices
They try to find the best value to replace the missing data	We take into consideration the group of nodes repotting data	We consider the proximity of: - the location - the reported data
The detection of the exact distribution of data can be difficult	We adopt a spatio-temporal approach	Our two-layer clustering scheme combined with a consensus model assists in the missing values replacement

The Envisioned Setting

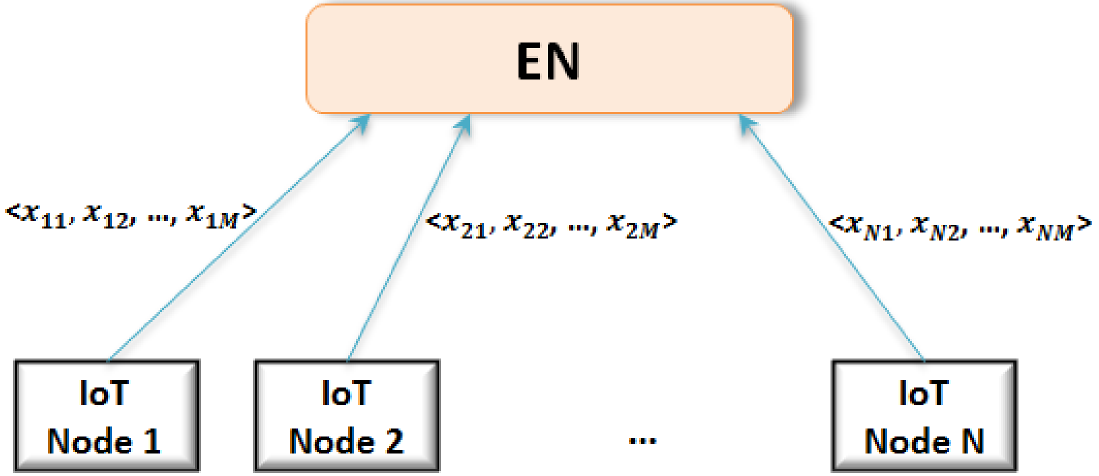
IoT devices collect multivariate data from their environment

A set of Edge Nodes (ENs) receive data from a set of IoT devices

Every EN should detect if missing values are present and, then, apply the proposed mechanism

Missing can refer in the entire multivariate vector or specific dimensions

We consider a sliding window approach and take into consideration the location of each device



Clustering and Correlation

1st level of clustering

- We focus on the spatial proximity of the devices
- We adopt the k-Means algorithm
- We create a set of clusters with devices in close distance
- Clusters are produced at pre-defined intervals being aligned with the mobility of the devices

2nd level of clustering

- We focus on the data proximity of the IoT devices
- The clustering is applied for a 1st level cluster
- We create a set of clusters with devices reporting ‘similar’ data
- We adopt the k-Means algorithm

Data Proximity

We get the IDs of the IoT devices and focus at each time step

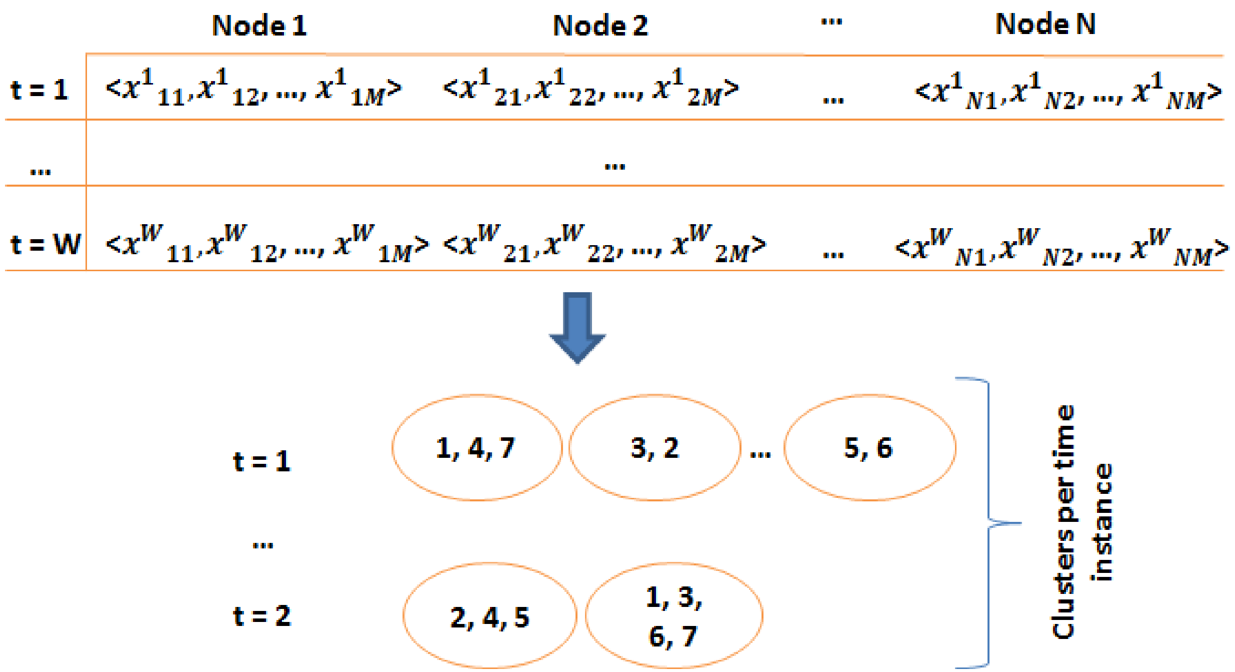
Every cluster represents a 'transaction' where a (sub-)set of IDs are involved

For each t, we provide the delivered clusters

The presence of IDs in a cluster at t depicts the data correlation between the corresponding IoT devices

When a missing value is present, we get the intersection of clusters where the device ID with missing value is present

We adopt the Pearson Correlation Coefficient (PCC) in multiple data dimensions



Data Imputation

Devices

We rely on the delivered clusters

We focus on the dimension where missing values are observed

When multiple dimensions are involved, we adopt an iterative approach

Aggregation

We adopt the linear opinion pool model

It is standard approach for aggregating multiple experts opinion

Only devices with strong correlation are involved

Imputation

We consider the weighted average of data at the same dimension

The weight of a device is high when a high correlation is observed

Weights are calculated on the correlation of all dimensions to avoid random events

Experimental Setup

We report on the performance of the model

We aim at detecting if the replacements are close to the real values

We adopt widely known performance metrics

Mean Absolute Error (MAE)

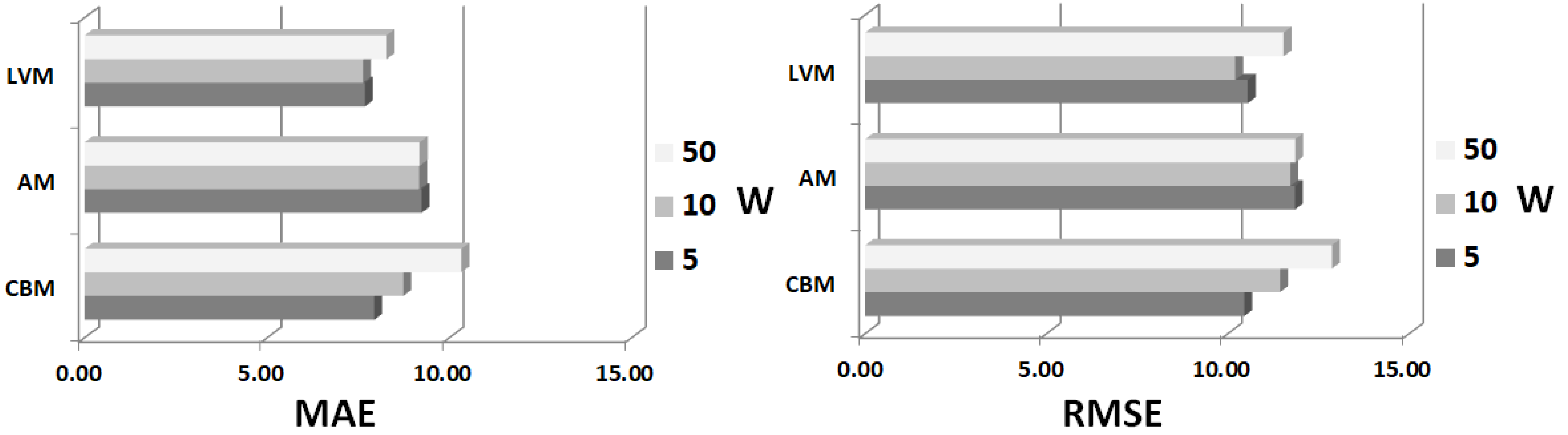
Root Mean Squared Error (RMSE)

Datasets:
- Unmanned Surface Vehicles Sensor Data[1]
- Intel Berkeley Research Lab Dataset[2]
- the Iris dataset[3]

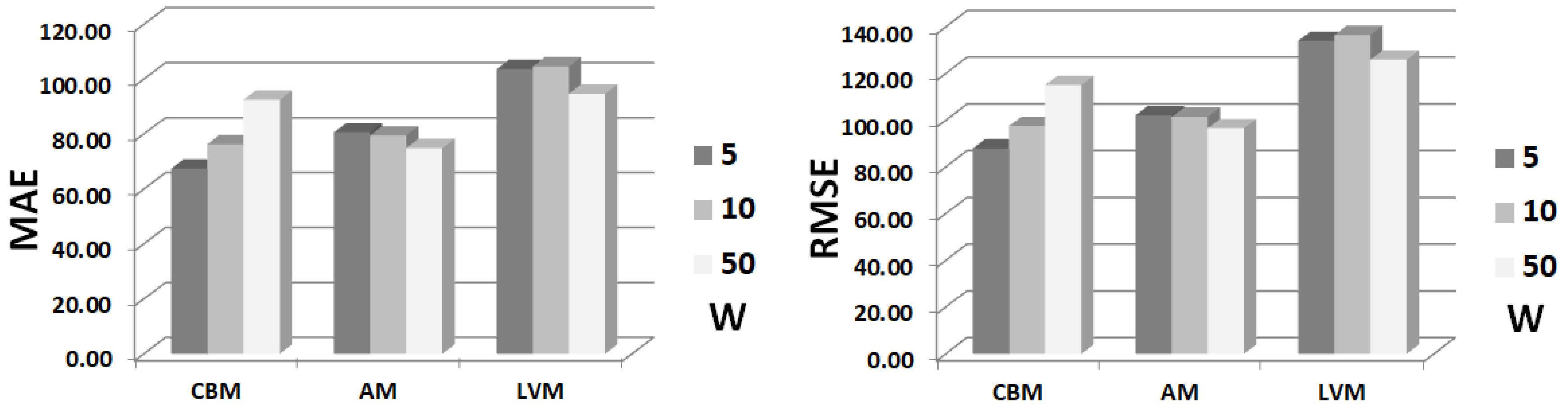
We compare our model with:
- an Averaging Mechanism
- the Last Value Mechanism

[1] Harth, N., Anagnostopoulos, C., 'Edge-centric Efficient Regression Analytics', IEEE EDGE, 2018
[2] http://db.csail.mit.edu/labdata/labdata.html
[3] http://archive.ics.uci.edu/ml/datasets/iris

Results

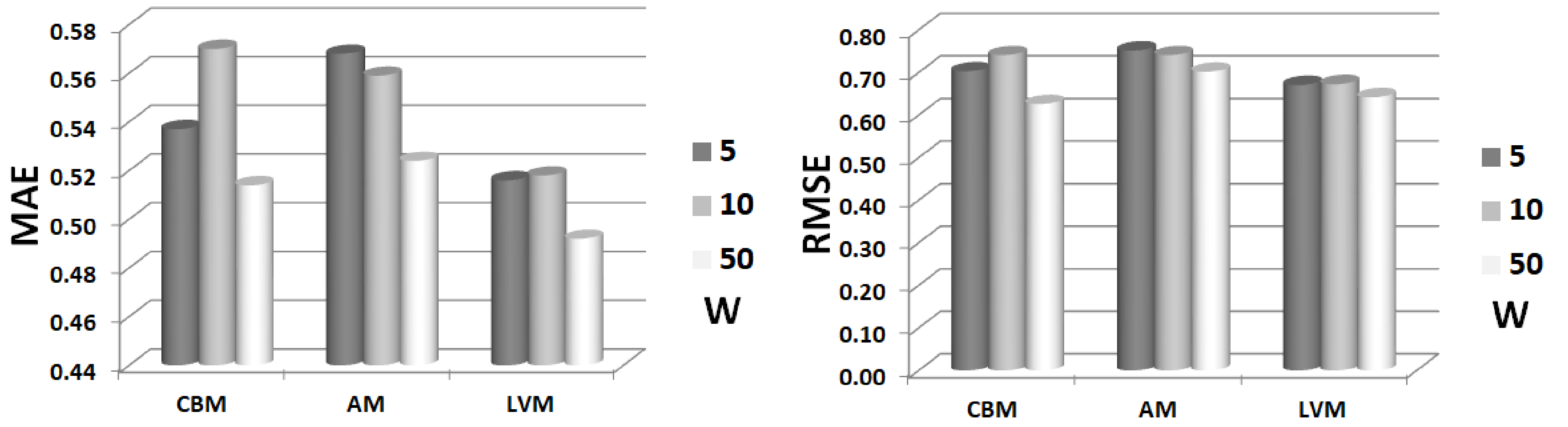


MAE and RMSE for the Unmanned vehicles dataset (different window values)

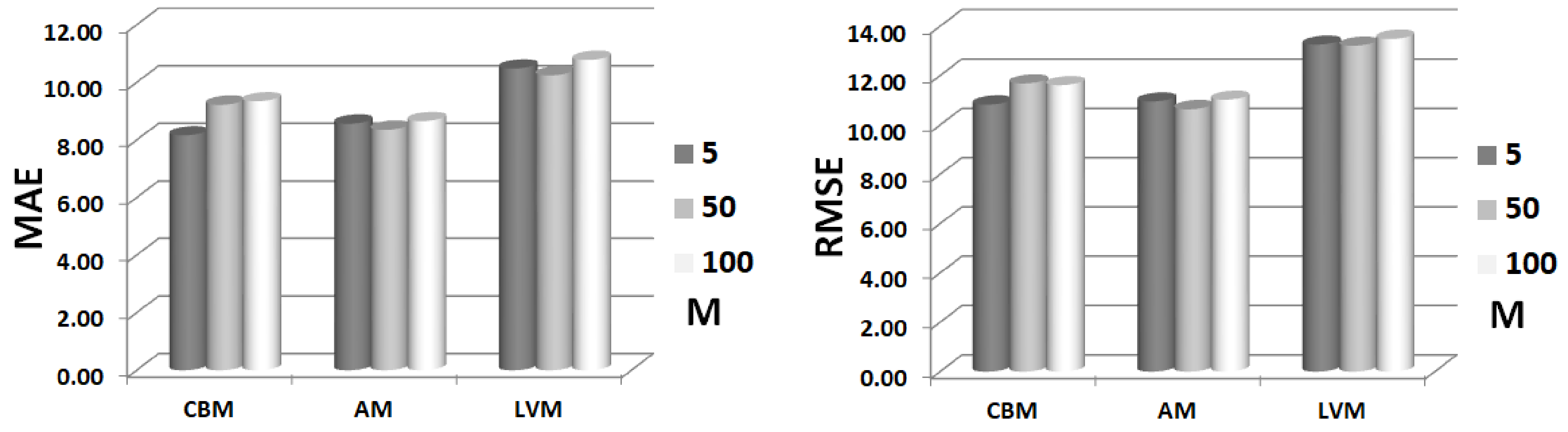


MAE and RMSE for the Intel dataset (different window values)

Results



MAE and RMSE for the Iris dataset (different window values)



MAE and RMSE for the Unmanned vehicles dataset (different number of dimensions)

Conclusions & Future Work

Efficiency

The proposed scheme can efficiently replace the missing values
Our mechanism outperforms in the comparative assessment for the majority of the experimental scenarios

Uncertainty

In the first places of our future research agenda is the management of the uncertainty in the aggregation process

Realization

The performance is better when the number of dimensions is low
The model is affected by the clustering process and the number of dimensions involved in the calculations

Statistics

Another research plan is to combine legacy techniques with the proposed model

Thank You!

Questions?