UNIVERSITY OF THESSALY

Intelligent Pervasive Systems Research Group
Dept. of Informatics and Telecommunications

Proactive, Correlation Based Anomaly Detection at the Edge.

Panagiotis Fountas, Kostas Kolomvatsos
Email: pfountas@uth.gr, kostasks@uth.gr

ICTAI 2021
November 1-3,2021
33rd IEEE International Conference on Tools with Artificial Intelligence.

Outline

- Introduction
- Problem Description
- The Ensemble Scheme
- Experimental Evaluation
- Conclusions & Future Work

Introduction

Numerous Devices

The increased adoption of Internet of Things (IoT).
The development of IoT application and usage IoT devices.

Huge Volumes of Data

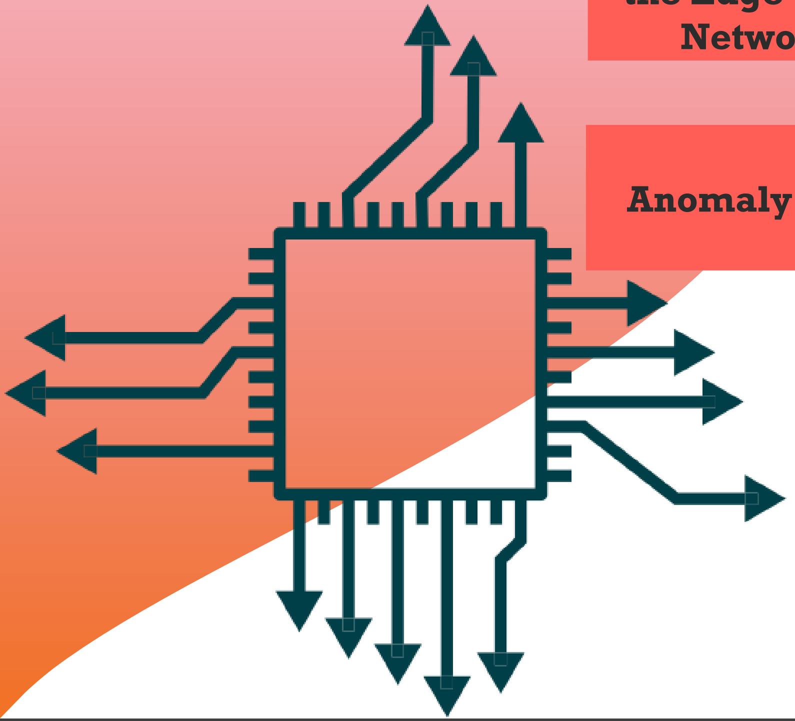
IoT devices and applications produce or collect data.
The data processed to create knowledge.

Interaction with the Edge of the Network

The Edge Computing (EC) nodes are placed close to the data sources.
Edge minimizes the latency in the provision of responses.
Edge Computing perform analytics over distributed data streams.

Anomaly Data

Data streams which differentiate from the distribution of the remaining data.
The goal is to detect and remove anomalies improving the performance of the processing activities.



Problem Description

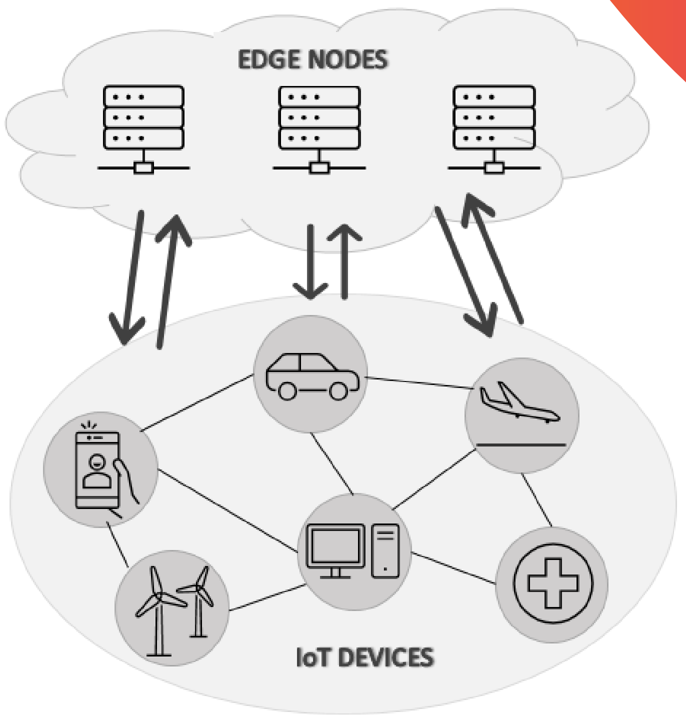
IoT devices/applications collect multivariate data from their environment.

Edge Nodes host geo-distributed datasets that consist of the reports of the IoT devices.

Edge Nodes are responsible to perform pre-processing activities

Edge Nodes use the proposed ensemble scheme to detect anomalies in the collected data.

Time instance	1st dimension	2nd dimension	…	d^{th} dimension
0	$r_1^j[0]$	$r_2^j[0]$	…	$r_d^j[0]$
1	$r_1^j[1]$	$r_2^j[1]$	…	$r_d^j[1]$
…	…	…	…	…
W	$r_1^j[W]$	$r_2^j[W]$	…	$r_d^j[W]$



The Ensemble Scheme

01

We adopt an extendable sliding window approach to focus only on the most recent data vectors.

02

We focus on the estimation of the correlation between two types of groups:

i. most correlated devices based on historical measurements (Model A),

ii. the nearest peers based on the most recent report (Model B).

03

Each model relies on the top-k peers which detected for the corresponding model i.e., $T_j^m = \{I_v^1, I_v^2, \dots, I_v^k\}$ where $m = \{A, B\}$ for Models A & B, respectively.

04

The proposed mechanism creates for each device a new target dataset $D_j^m = \{Q_j^1 \cup Q_v^1 \cup \cdots \cup Q_v^k\}$ based on the sub-dataset of each device i.e., $Q_j = \{R_0^j, \dots, R_{W-1}^j\}$ and of the top-k peers depicted for each model.

05

The modified DBSCAN is applied over the target datasets which are created for each device i.e., D_j^m.

06

Both models (A & B) generate the corresponding anomalies estimation list as follows:

$P_m^y = \{p_{m_0}^1, \dots, p_{m_{w-1}}^1, \dots, p_{m_0}^N, \dots, p_{m_{w-1}}^N\}$ where y depicts the window, $p_{m_t}^j = \{R_t^j : pr_{val}\}$ and $pr_{val} = \{-1,1\}$.

The Ensemble Scheme

07

The proposed mechanism is based on the lists P_A^y, P_B^y and the final estimation is built upon the aggregation of the two lists, i.e., $\mathrm{Fp_y} = \{fp_0^1, \dots, fp_{w-1}^1, \dots, fp_0^N, \dots, fp_{w-1}^N\}$

08

The final estimation is an object which has two attributes:

i. $pr_{val_t}^j = \{-1,0,1\}$ where -1:Outlier, 0:Potential outlier and 1:Inlier

ii. label $l_{fp_t}^j = \{Conf_{anomaly}, Pot_{anomaly}, Inlier\}$.

09

The realization of attributes is dictated by the following rules:

- If pr_{val} of $p_{A_t}^j$ and $p_{B_t}^j$ for R_t^j are the same and equal to -1, then the object's attributes take the following values $fp_t^j = \left\{pr_{val_t}^j = -1, l_{fp_t}^j : Conf_{anomaly}\right\}$
- If pr_{val} of $p_{A_t}^j$ and $p_{B_t}^j$ for R_t^j are the same and equal to 1, then the object's attributes take the following values $fp_t^j = \left\{pr_{val_t}^j = 1, l_{fp_t}^j : Inlier\right\}$
- If pr_{val} of $p_{A_t}^j$ and $p_{B_t}^j$ for R_t^j differ then the object's attributes take the following values $fp_t^j = \left\{pr_{val_t}^j = 0, l_{fp_t}^j : Pot_{anomaly}\right\}$

10

Potential anomalies are placed in a separate list for further investigation.

11

We argue on investigating potential anomalies by incorporating more data into our reasoning to confirm our final decision.

We slightly increase W by a factor of $ex = \frac{W}{3}$.

Hence, we can perform our processing for a new window $W' = W + ex$ with additional data.

The Ensemble Scheme

12

We fire again the Models A & B and get the corresponding estimations for the new window.
$P'^{y}_{A} = \left\{p^{1}_{A_0}, \dots, p^{1}_{A_{w'-1}}, \dots, p^{N}_{A_0}, \dots, p^{N}_{A_{w'-1}}\right\}$ and $P'^{y}_{B} = \left\{p^{1}_{B_0}, \dots, p^{1}_{B_{w'-1}}, \dots, p^{N}_{B_0}, \dots, p^{N}_{B_{w'-1}}\right\}$.

13

Using P'^{y}_{A} and P'^{y}_{B} and based on the rules of the previous slide, we produce the final estimation list for the extended window i.e., $\mathrm{Fp'}_{\mathrm{y}} = \{fp'^{1}_{0}, \dots, fp'^{1}_{w-1}, \dots, fp'^{N}_{0}, \dots, fp'^{N}_{w-1}\}$.

14

The proposed mechanism draws the final estimations for the potential anomalies detected in the previous phase upon W data vectors) using the following rules:

- If $l'^{j}_{fp'_t}$ is $Conf_{anomaly}$, then the estimation for the fp^{j}_{t} is updated to $fp^{j}_{t} = \left\{pr^{j}_{val_t} = -1, l^{j}_{fp_t}: Conf_{anomaly}\right\}$
- If $l'^{j}_{fp'_t}$ is $Pot_{anomaly}$, then the estimation for the fp^{j}_{t} is updated to $fp^{j}_{t} = \left\{pr^{j}_{val_t} = -1, l^{j}_{fp_t}: Conf_{anomaly}\right\}$
- If $l'^{j}_{fp'_t}$ is $Inlier$ then the estimation for the fp^{j}_{t} is updated to $fp^{j}_{t} = \left\{pr^{j}_{val_t} = 1, l^{j}_{fp_t}: Inlier\right\}$

Experimental Evaluation

Parameters
Number of top-k correlated/closest devices $k \in \{2,3,4\}$
Percentage of anomalies in dataset $V = 5\%$
Number Nodes in network $\mathrm{N} = 5$
Sliding Window size $W = 163$
Neighborhood threshold $T = \{0.995, 0.996, 0.997, 0.998, 0.999\}$

Dataset	Source
Greenhouse dataset	http://www.iprism.eu/assets/greenhouse_dataset_sept_2020.csv

Precision= $\frac{TP}{TP+FP}$

1st Metric

Recall= $\frac{TP}{TP+FN}$

2nd Metric

Accuracy= $\frac{TP+TN}{TP+TN+FP+FN}$

3rd Metric

TNR= $\frac{TN}{TN+FP}$

4th Metric

Experimental Evaluation

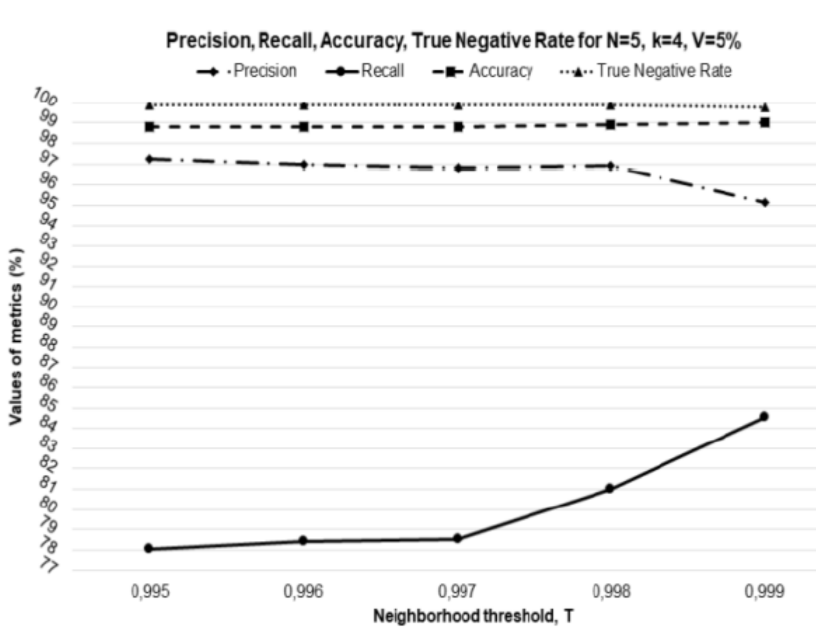


Figure 1

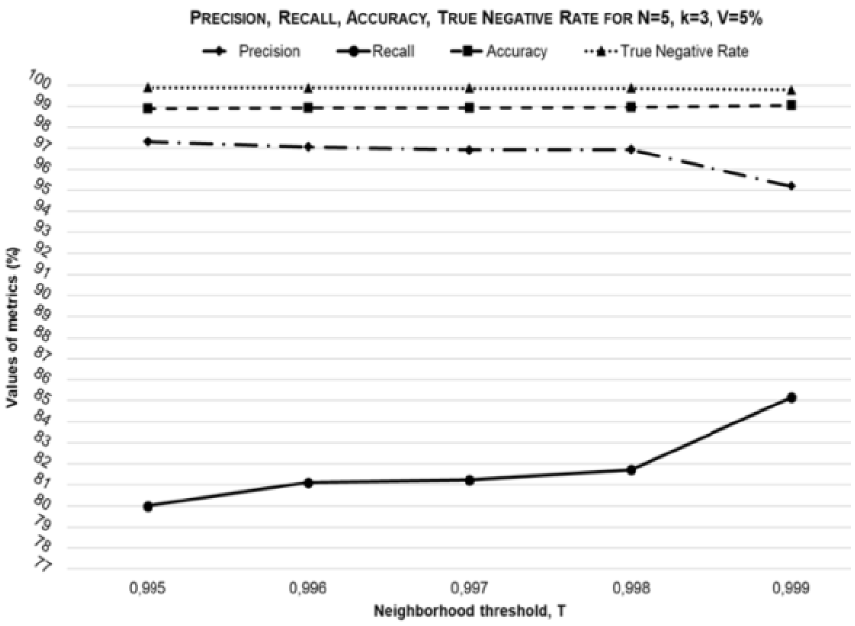


Figure 2

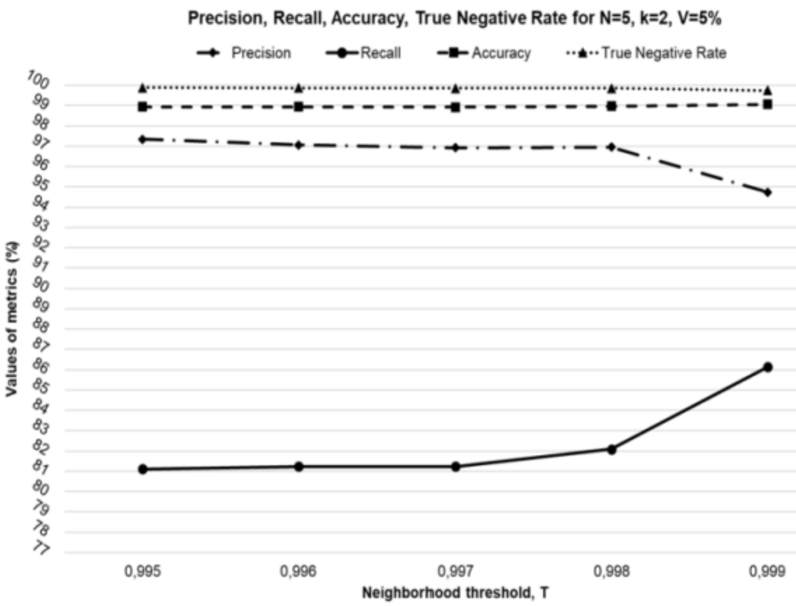


Figure 3

- We test the performance of the model through the values of variable k and T:
  - Precision and TNR are negatively affected by the increment of T in all experimental scenarios.
  - Recall and Accuracy are positively affected by the increment of T in all experimental scenarios.
  - The increment of k does not have impact on Accuracy and TNR while it has an impact on Recall and Precision in opposite directions.
- We conclude that the decrease of k in combination of the extension of the window size when there is necessary clearly affect the performance of our model

Conclusions & Future Work

- Anomaly detection at the edge is a significant research subject.
- The processing activities can be more efficient through the detection and removal of anomaly data.
- It is necessary to provide models, algorithms, and techniques that are capable to detect anomaly data with a high accuracy.
- Our future research plans involve more complex models for the management of the sliding window.

iPRISM

THANK YOU

Panagiotis Fountas
Email: pfountas@uth.gr

http://www.iprism.eu