UNIVERSITY OF THESSALY
creative years

Intelligent Pervasive Systems Research Group
Dept. of Informatics and Telecommunications

Landmark based Outliers Detection in Pervasive Applications

Kostas Kolomvatsos
kostasks@uth.gr

12th International Conference on Information and Communication Systems (ICICS 2021)
24-26 May 2021 Valencia - Spain (Virtual)

OUTLINE

- Introduction
- Problem Description
- Magnitude of Outliers
- Detection Strategy
- Experimental Evaluation
- Conclusions & Future Work

INTRODUCTION

Numerous Devices	The increased adoption of the 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 should have appropriate form to draw conclusions
Interaction with the Edge of the Network	The Edge Computing (EC) nodes are placed close to the data sources. Edge minimize the latency in the provision of responses. Edge Computing perform analytics over distributed data streams.
Processing Activities	Data processing based on the requested tasks Simple or more complex processing for delivering statistical information of datasets.

CONTRIBUTION

- We study a novel model for outliers detection departing from the state of the art
- The vast majority of the relevant efforts in the domain adopt a ‘one-shot’ decision making
- We focus on a mechanism that applies tolerance in the outliers detection process
- Every outlier data is not directly confirmed as an anomaly but we apply a temporal management to deliver a set of candidate outliers
- Candidates are confirmed upon the new data that arrive into the system
- The confirmation of outliers is based on a landmark window expanded to incorporate more data into our process

PROBLEM DESCRIPTION

- We consider a set of edge nodes that are owners of distributed datasets
- Contextual data vectors are reported by IoT devices that capture them through interaction with their environment and users
- We rely on a combination of a sliding window and a landmark window approach
- We identify potential outliers in the last W observations (sliding window)
- These are the ‘candidate’ outliers annotated for further investigation
- We alter our processing and adopt a landmark window to incorporate more data objects
- The maximum size of the landmark window is (at most) twice the sliding window
- We re-evaluate candidate outliers and their new status
- Confirmed outliers are evicted from the dataset

MAGNITUDE OF OUTLIERS

- We define the concept of the magnitude of the candidate outlier based on its distance to the population (the dataset) <u>calculated over the mean of the k highest distances</u>

- We consider that the 'fuzzy' notion of the magnitude of each outlier is measured by a sigmoid function (x: distance, α, β: smoothing parameters)

$$\lambda = \frac{1}{1+e^{(-\alpha x+\beta)}}$$

- When the distance exceeds a threshold (as defined by the realization of the aforementioned sigmoid function), the magnitude of the outlier indication is very high (close to unity)

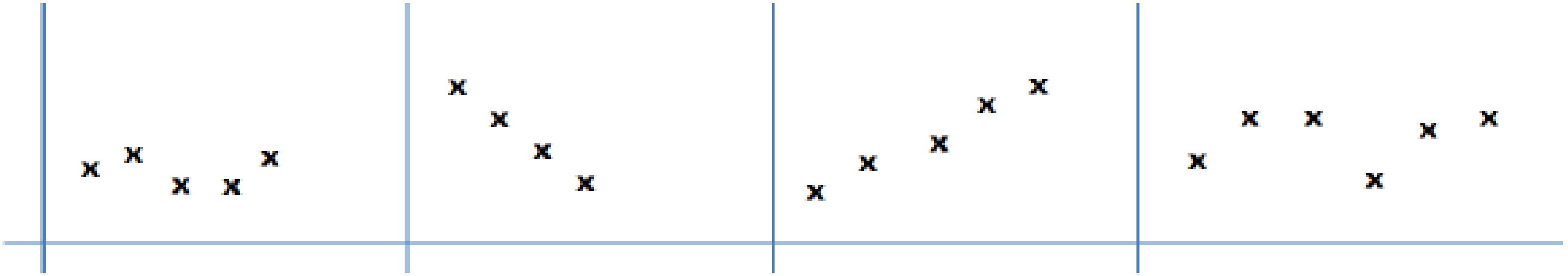
- When the adoption of the landmark window is decided, we increase the size of the window by adding a small amount of discrete time instances

- Then, we record and monitor the realization of λ – its trend plays a significant role in the confirmation of outliers

- We perform again the calculations for exposing the distance of candidate outliers from the population

DETECTION STRATEGY

- Our non parametric trend analysis is applied upon λ realizations as the landmark window is expanded
- For trend analysis, we adopt an ensemble scheme upon the widely known Mann-Kendall metric or Mann-Kendall test (MKM) and the Sen's slope (SS)
- We rely on a simple and fast, however, efficient technique to aggregate both results
- The easy scenario is met when both techniques agree upon the trend of λ (increasing or decreasing)
- In case of an disagreement, we consider a 'strict' boolean model which relies on a conjunctive form
- Disagreements are solve by deciding a 'neutral' view for λ
- If the final outcome indicates an increasing trend, we consider the data as a confirmed outlier
- If the trend is detected as decreasing or neutral and the distance is below a threshold θ, the data are accepted as a normal value



EXPERIMENTAL EVALUATION

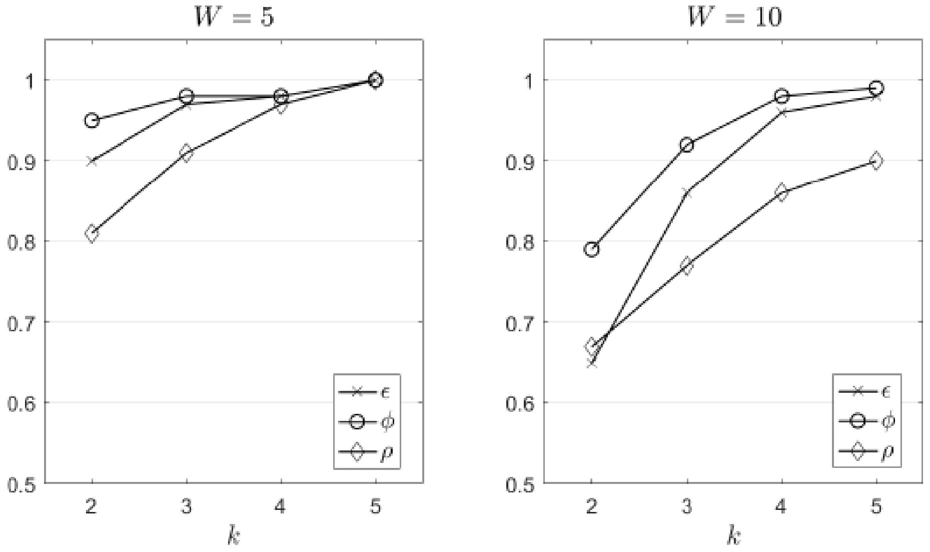
- We report on the experimental evaluation of the proposed model based on a custom simulator

- Performance metrics:
  - ε: **Accuracy** is defined as the number of correct detections out of the total number of the identified outliers
  - υ: **Precision** is defined as the fraction of the correctly detected outliers
  - r: **Recall** is defined as the fraction of the detected outliers that are successfully retrieved compared to the true outliers
  - ϕ: **F-measure** is a combination of υ and r defined as follows
  - ρ: the **area under the ROC curve** (ROC AUC), i.e., the mean of the recall (r) upon the top-ranked objects (in the list of the potential outliers)

- Two real datasets (references are provided in the paper)
  - The ionosphere dataset has 32 numeric dimensions and 351 instances where 126 outliers (35.9%) are detected
  - The Wisconsin Prognostic Breast Cancer (WPBC) dataset has 33 numerical dimensions and 198 instances where 47 outliers (23.74%) are detected

EXPERIMENTAL EVALUATION

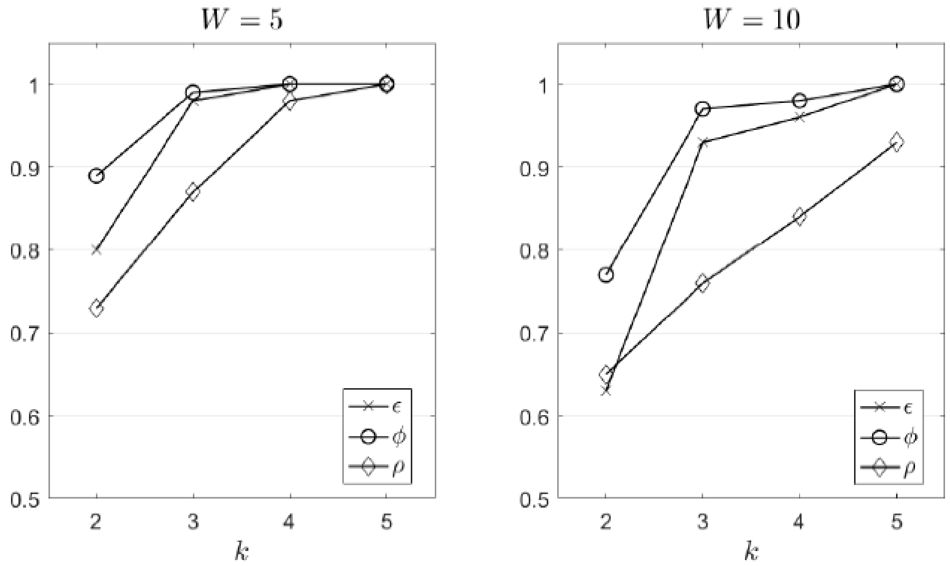
- We observe that an increased number of neighbours positively affects the performance of our model
- As k increases, the adopted metrics reach very close to the optimal value
- False positives and false negatives events are minimized
- When W = 10 (see Figure - right), we observe a similar performance, however, the outcomes are lower than in the previously presented experimental scenario (W=5)
- A sub-set of candidate outliers are not finally confirmed and are incorporated into the dataset



Performance outcomes for Dataset 1

EXPERIMENTAL EVALUATION

- The outcomes for Dataset 2 are similar as in the previous experimental scenario
- These evaluation results confirm our observations
- We observe that the proposed model clearly outperforms other efforts for various realizations of k
- Other relevant models, evaluated for the same dataset, achieve the maximum ϕ in the interval [0.38, 0.44] while ρ is realized in the interval [0.46, 0.58]



Performance outcomes for Dataset 2

CONCLUSIONS & FUTURE WORK

- ❖We propose the use of a model that, based on a 'soft' approach, decides the presence of outliers in a dataset
- ❖We define the concepts of candidate and confirmed outliers as well as the magnitude of the difference of an outlier from the remaining population
- ❖Our temporal management process builds upon the combination of a sliding with a landmark window
- ❖The proposed technique is experimentally evaluated and its advantages and disadvantages are revealed
- ❖Our future plans involve the adoption of a scheme based on Fuzzy Logic and machine learning to be able to expose more complex trends and connections between data objects

iPRISM

THANK YOU

Kostas Kolomvatsos
kostasks@uth.gr

http://www.iprism.eu