UNIVERSITY OF THESSALY
creative years

Intelligent Pervasive Systems Research Group
Dept. of Informatics and Telecommunications

Probabilistic Data Allocation in Pervasive Computing Applications

Dr Kostas Kolomvatsos

19th IEEE International Conference on Ubiquitous Compputing and Communications (IUCC)
December 17-19, 2020

Outline

- Introduction
- Problem Description
- Probabilistic Outliers Detection
- An Ensemble Mechanism for Outliers Detection
- Statistical Management of Data Vectors
- Experimental Evaluation
- Conclusions & Future Work

Introduction

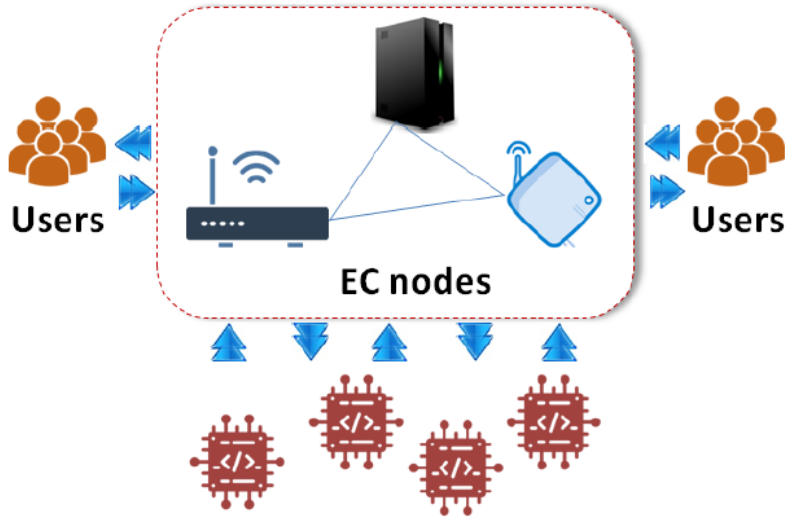
- Pervasive Computing (PC) deals with the placement of services and applications around end users for facilitating their everyday activities
- The combination of the Internet of Things (IoT) and Edge Computing (EC) provides a promising 'aggregation' of two different infrastructures for supporting innovative Pervasive Computing (PC) applications
- IoT devices are, usually, carried by end users or they are present in the environment in close distance with them facilitating the envisioned interactions and the collection of data
- Data become the subject of processing activities to create knowledge and support novel applications
- Edge nodes have direct connection with IoT devices and become the hosts of a high number of geo-distributed datasets
- A challenge is to keep the consistency and accuracy of data at high levels as the critical statistical information that depicts the quality of the collected data

Contribution

- We propose a model for concluding solid datasets, i.e., datasets that exhibit a high accuracy realized when the error/difference between the involved data is low (e.g., the standard deviation may be limited)
- We also focus on a data replication approach to support a fault tolerant EC infrastructure
- Combining a replication activity with distributed data storage, we are able to reduce the probability of data loss and efficiently manage failures of nodes
- We provide an ensemble model for the aggregation of multiple outlier indicators upon a double majority voting scheme
- We support the data replication process with a probabilistic model based on multiple historical synopses reported by edge nodes
- Our aim is to detect the similarity of the incoming data with the available datasets upon their past trends

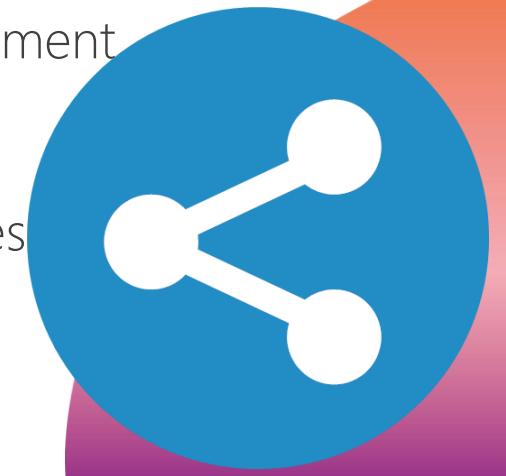
Problem Description

- We consider a set of edge nodes that are owners of distributed datasets (DS)
- Contextual data vectors are reported by IoT devices that capture them through interaction with their environment and users
- We consider the online knowledge extraction model as the statistical synopsis for each dimension of the multivariate data
- Synopses can be useful when we want to have a view on the collected data in a remote location
- Edge nodes try to act in a cooperative manner and decide to exchange their synopses regularly



Problem Description

- Initially, every node should detect if the incoming data vector is an outlier not only compared with the local dataset but also with the remaining repositories present at peer nodes (with the assistance of data synopses)
- Our aim is to detect if data significantly deviate from the 'ecosystem' of datasets, thus, no dataset could become the host of x
- We rely on an ensemble approach and study the involvement of multiple outlier detection methods
- The incoming data are rejected when multiple indicators depict an outlier judgment in multiple datasets
- When data are not outliers, they are stored locally and replicated to peer nodes exhibiting a significant correlation



Probabilistic Outliers Detection

- A set of outlier indicators for each of the available datasets feed a two-dimensional matrix I
- In general, I is a matrix hosting the outcomes of V X N Bernoulli trials (1: outlier, 0: non outlier) with different success probabilities (V is the number of the detectors)
- Our aim, before we decide that a data vector is an outlier, is to detect if multiple indicators agree upon this event for multiple datasets
- We perform a double majority voting, i.e., the first per column (multiple indicators for the same dataset) and the second upon multiple aggregated indicators for the ecosystem of datasets
- We rely on the δ-majority function upon V binary variables (δ is the threshold over which we consider a data vector as an outlier for a specific dataset)

	DS_1	DS_2	DS_3	…
V_1	1	0	1	
V_2	0	0	0	
…				

Probabilistic Outliers Detection

- As I is the set of V X N Bernoulli trials with different success probabilities, we can easily deliver the final success probability for any incoming data vector
- The sum of the outcomes of the aforementioned Bernoulli trials can be adopted to define the variable Z which follows a Poisson Binomial distribution
- We can easily define the success probability that a data vector will be identified as outlier in the entire group of detectors and nodes under the principle of majority voting (N is the number of nodes/datasets and β is the individual success probability)

$$F(z) = \sum_{m=z}^{N} \left\{ \sum_{A \in \mathcal{F}_m} \prod_{i \in A} \beta_{ij} \prod_{j \in A^c} (1 - \beta_{ij}) \right\}$$

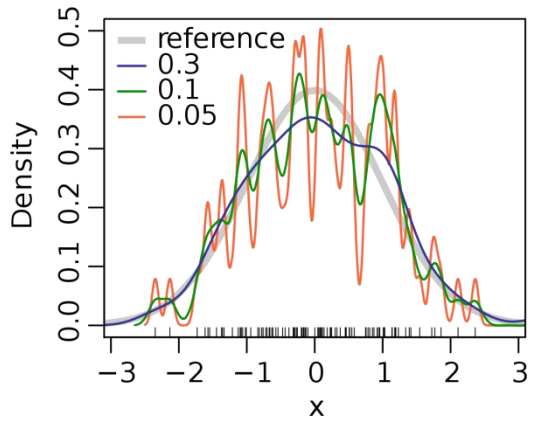
$$z = \begin{cases} \frac{N}{2} + 1 & \text{if } N \text{ is even} \\ \frac{N+1}{2} & \text{if } N \text{ is odd} \end{cases}$$

Statistical Management of Data Vectors

- The replication process refers in nodes that are highly correlated with the non-outlier data
- Our replication process is realized upon the collected synopses
- We detect the trend of the correlation between the incoming data and the available synopses
- We rely on the similarity between the data vector and each synopsis retrieved by peer nodes
- Based on the collected synopses, we can have a time series of distances with data
- Upon these distances, we can expose their unknown pdf targeting to extract the probability of having the similarity upon a pre-defined threshold

Statistical Management of Data Vectors

- For estimating the unknown pdf, we adopt the Kernel Density Estimator with the Gaussian as the kernel function
- We get the probability of having the similarity between a data vector and a dataset upon a pre-defined threshold
- We also rely on the Probability Ranking Principle which dictates that if peers are ordered by decreasing order of the calculated probability, the model is the best to be gotten for those instances
- From the ranked list, we select the top-k outcomes and the corresponding nodes to host the incoming data vector



Experimental Evaluation

- We investigate the performance of our mechanism concerning the number of the detected outliers
- We investigate the number of the stored data vectors that deviate from the statistics of every dataset
- We also focus on the evaluation of the proposed model concerning its ability to keep datasets solidity at high levels

- Performance metrics:
  - the percentage of the detected outliers ω (we randomly 'produce' fake outliers)
  - the percentage of data vectors that deviate from the statistics of each dataset τ
  - the solidity of the formulated datasets as depicted by the mean μ and standard deviation σ

- Dataset: a real dataset with 9,358 instances of hourly averaged responses from an array of five (5) metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device

Experimental Evaluation

- The proposed model is capable of detecting the generated outliers (ω in [0.70, 1.0]) while keeping similar data to the same datasets especially when N is low (W is the window of values taken into consideration in our calculations – k is the number of nodes where data vectors are replicated)
- A high number of nodes N leads the dispersion of datasets to increase
- This situation is also affected by the increased number of nodes selected to replicate every accepted data vector (k = 5)

PERFORMANCE OUTCOMES FOR VARIOUS EXPERIMENTAL SCENARIOS AND $W = 10$

	$k=2$				$k=5$			
	$M=2$		$M=10$		$M=2$		$M=10$	
N	ω	τ	ω	τ	ω	τ	ω	τ
10	1.00	0.005	0.90	0.10	1.00	0.005	0.85	0.11
50	1.00	0.08	1.00	0.28	1.00	0.10	1.00	0.30
100	1.00	0.10	1.00	0.30	1.00	0.20	1.00	0.40

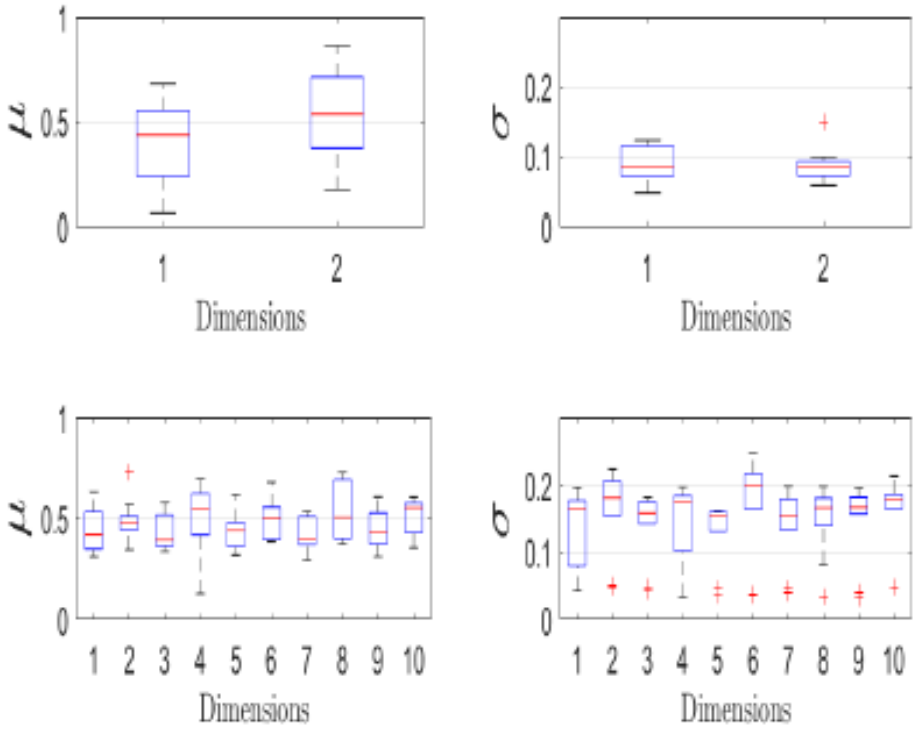
PERFORMANCE OUTCOMES FOR VARIOUS EXPERIMENTAL SCENARIOS AND $W = 50$

	$k=2$				$k=5$			
	$M=2$		$M=10$		$M=2$		$M=10$	
N	ω	τ	ω	τ	ω	τ	ω	τ
10	0.87	0.002	0.70	0.004	0.95	0.05	0.70	0.02
50	1.00	0.10	1.00	0.10	1.00	0.06	1.00	0.12
100	1.00	0.20	1.00	0.30	1.00	0.18	1.00	0.33

Experimental Evaluation

- The number of dimensions M affects the dispersion of data (N=10) - We observe that σ is low when M is low as well

- The aggregation of the difference between the available synopses and data vectors when M is high may jeopardize the solidity of datasets



Conclusions and Future Work

- We propose a probabilistic ensemble scheme for outliers detection in a set of datasets
- We identify the datasets that match to the incoming vectors to support an efficient replication process
- Our decision making is applied upon historical synopses
- Our model is capable of concluding any allocation decision in the limited possible time, i.e., in real time
- The outliers detection rate is optimal for the vast majority of the experimental scenarios while the solidity of the formulated datasets is kept at high levels
- One of our future research plans is to incorporate a communication model between nodes and include the aspects of that model into the decision making

iPRISM

THANK YOU

Dr Kostas Kolomvatsos
kostasks@uth.gr
http://www.iprism.eu