UNIVERSITY OF THESSALY
creative years
1988 — 2018

Intelligent Pervasive Systems Research Group
Dept. of Informatics and Telecommunications

Proactive Data Allocation in Distributed Datasets based on an Ensemble Model

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
- Synopses Management
- The Ensemble Similarity Model
- Experimental Evaluation
- Conclusions & Future Work

INTRODUCTION

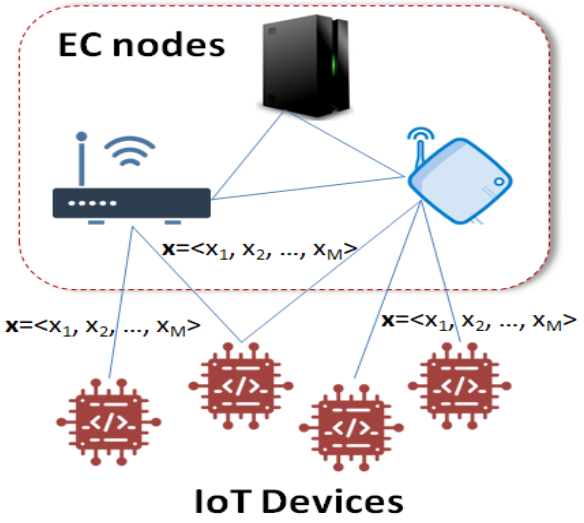
Numerous Devices	The increased adoption of Internet of Things (IoT). The development of IoT application and usage IoT devices.
Huge Volumes of Data	IoT devices and applications produce or collect data. The data should have appropriate form to draw conclusions
Interaction with the Edge of the Network	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.
Data Management	Data separation for easy management Data allocation at the appropriate datasets to secure the accuracy and similarity Create a 'map' of the available data

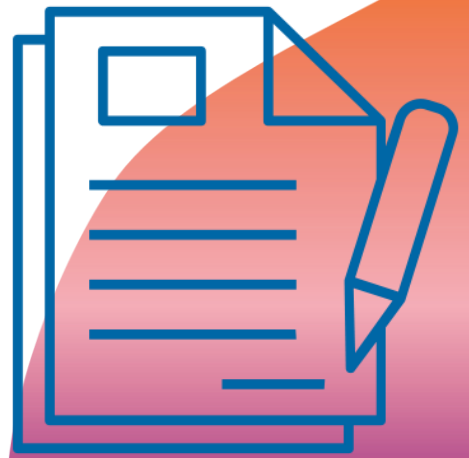
CONTRIBUTION

- We act in a proactive manner and reason over the best possible action towards the minimization of the error in the available data partitions
- We decide to deliver the similarity between the incoming data and the available partitions through the adoption of synopses
- We do not have to process the entire dataset when new data arrive and propose an incremental model for updating synopses
- We adopt an ensemble scheme upon the statistical similarity between data and the available partitions
- Our method also supports the provision of partitions with the minimum overlapping as data are placed at the datasets

PROBLEM DESCRIPTION

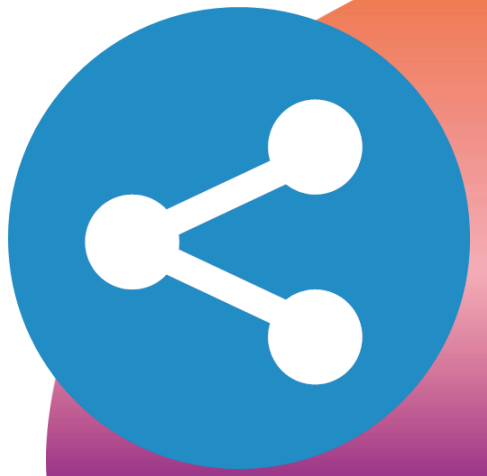
- We consider a set of N edge nodes that are owners of multivariate distributed datasets $D=\{D_1, D_2, \ldots, D_N\}$
- The corresponding synopses $S=[s_1, s_2, \ldots]$ are calculated upon the stored data vectors $D_i=[x1, x2, x3, \ldots]$
- Synopses become the basis for deciding where the incoming data will be allocated
- We adopt an ensemble scheme for that, i.e., we try to find where the similarity between a data vector and the available synopses is maximized
- The updates of synopses can be performed in an incremental manner, i.e., when new data arrive, synopses are realized as the 'extension' of the previous calculated version and the new data
- Nodes receiving N synopses and data vectors apply the proposed ensemble scheme and detect the appropriate dataset to host the data





SYNOPSES MANAGEMENT

- We rely on a fast technique as, especially the update process, should be incrementally realized in the minimum time
- We consider that an online micro-clustering algorithm is adopted for delivering the necessary synopses
- The idea is adopted by the BIRCH algorithm
- When new data arrive in a node, the algorithm finds the closest cluster and, accordingly, it updates the leafs and the internal nodes
- We provide a module that scans the CF-tree and finds the delivered clusters that contain at least α data vectors
- We focus on the part of data that dominates the dataset/partition and exclude data that are not similar with the majority



THE ENSEMBLE SIMILARITY MODEL

- We adopt three metrics that can be applied for positive numbers, i.e., Jaccard, Sorensen and Kulczynski metrics
- The outcome of these metrics are met in the interval [0,1]
- Jaccard dissimilarity is the proportion of the combined abundance that is not shared and defined as follows:

$$O_1 = \frac{2\sum_{j=1}^{M} x_{ij} - s_{hj}}{\sum_{j=1}^{M} x_{ij} + \sum_{j=1}^{M} s_{hj} + \sum_{j=1}^{M} x_{ij} - s_{hj}}$$

- Sorensen metric is also known as the Bray-Curtis coefficient and targets to the detection of the shared abundance divided by the total abundance

$$O_2 = \frac{\sum_{j=1}^{M} x_{ij} - s_{hj}}{\sum_{j=1}^{M} x_{ij} + \sum_{j=1}^{M} s_{hj}}$$

- The Kulczynski metric measures the arithmetic mean probability that if one object has an attribute, the other object has it too

$$O_3 = 1 - \frac{1}{2}\left[\frac{\sum_{j=1}^{M} \min(x_{ij}, s_{hj})}{\sum_{j=1}^{M} x_{ij}} + \frac{\sum_{j=1}^{M} \min(x_{ij}, s_{hj})}{\sum_{j=1}^{M} s_{hj}}\right]$$

THE ENSEMBLE SIMILARITY MODEL

- We rely a linear opinion pool as the aggregation function
- We adopt the aggregation opinion operator, i.e., $O' = g(O_1, O_2, \ldots, O_{|O|}) = w_1 O_1 + w_2 O_2, \ldots + w_{|O|} O_{|O|}$ where w_i is the weight associated with the measurement of the ith metric's outcome O_i such that $w_i \in [0,1]$ and $w_1 + \ldots + w_{|O|} = 1$
- Weights are calculated based on specific characteristics that affect the confidence on each similarity outcome
- We adopt a simple outliers detection technique based on the statistics of the outcomes
- We consider that if a result deviates for more than three times the deviation of metrics' outcomes from the mean of the outcomes is considered as an outlier (the adoption of the Gaussian distribution is an assumption towards the discussed target).
- When a result is detected as an outlier, the specific metric gets a very low weight θ (e.g., $\theta=0.1$) and the remaining metrics equally share the difference form the unity $1-\theta$

EXPERIMENTAL EVALUATION

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

- A real dataset (references are provided in the paper)
  - The air quality dataset with 9,358 instances of hourly averaged values from five (5) metal oxide chemical sensors being embedded in an air quality multisensor device
  - Initially, we separate the data into a set of datasets (e.g., 5) by randomly selecting instances and adopt five dimensions (from those defined in the original dataset)

- Data vectors are produced based on a specific mean (μ) and standard deviation (σ) before we match them against the synopses calculated over the available separated datasets

- We adopt three (3) experimental scenarios as follows: (i) μ=25, σ=10; (ii) μ=25, σ=20; (iii) μ=50, σ=50

- We pay attention on the mean and the standard deviation for each dimension after using the proposed model and placing the incoming data into the most similar dataset/partition

EXPERIMENTAL EVALUATION

- We observe that the mean of the adopted dimensions are very close exposing the ability of the model to collect similar values into the same datasets
- The resulted partitions exhibit a standard deviation lower than the deviation adopted to produce the data vectors
- The realizations of the mean and the deviation are close for all the involved dimensions

SUMMARY OF THE PRESENTED PERFORMANCE OUTCOMES.

Scenario	Data production		'majority' dataset	μ interval	σ interval
	μ	σ			
1st	25	10	4368	12,49-13,89	8,39 - 9,12
2nd	25	20	3947	13,90-14,00	11,07 -11,57
3rd	50	50	4782	28,69 - 29,53	25,83 -26,44

CONCLUSIONS & FUTURE WORK

- The allocation of the collected data to the appropriate datasets is significant for any future processing
- We can save resources and time if we 'pre-process' the data just after their arrival instead of performing a batch oriented approach upon huge volumes
- We propose a methodology for allocating the data to the most appropriate dataset from those that are available at the EC infrastructure
- We perform a set of experimental evaluations and reveal the ability of the proposed scheme to gather similar data to the same dataset
- Future improvements of the approach involve the adoption of a more complex ensemble scheme that will rely on the historical 'behavior' of similarity metrics

iPRISM

THANK YOU

Kostas Kolomvatsos
kostasks@uth.gr

http://www.iprism.eu