An Ensemble Interpretable Machine Learning Scheme for Securing Data Quality at the Edge

Anna Karanika[1], Panagiotis Oikonomou[1], Kostas Kolomvatsos[1], Christos Anagnostopoulos[2]

[1] Dept. of Informatics and Telecommunications. University of Thessaly, Lamia, Greece

[2] School of Computing Science. University of Glasgow, Glasgow, UK

Introduction

- Nowadays we are witnessing the advent of Internet of Things
  - humongous volumes of data
- Perform processing at the edge of the network
  - heterogeneous nodes
  - Close to IoT devices / end users
- Data collection is a key aspect of Edge Computing (EC) nodes
  - multivariate data
  - Data validation

Motivation

- Data quality is significant for any application
- Secure data quality at the edge
  - Accuracy and separation algorithms
- The 'curse' of dimensionality demands new solutions
  - features ↑ → samples ↑
  - poor performance of ML models
- Avoid over fitting
- Data quality and integrity

Motivation

- Our goal:

  is to provide a decision making model for securing data quality based on an ML scheme that will produce the relevant knowledge about the domain relationships

ensemble scheme

1. Permutation Feature Importance (PFI)
2. Shapley Values
3. Feature Interaction Technique (FIT)

Artificial Neural Network (ANN)

Secure data quality

Contributions

1) Prepare the data before the actual processing is applied
2) Interpretable ML scheme

   for satisfying the meaningful knowledge extraction
3) Ensemble scheme

   for aggregating multiple interpretable ML
4) Artificial Neural Network

   for judging the significance of every feature
5) Data exclusion

   that may lead to an increased error
6) minimum overlapping of the available datasets

System Model (1)

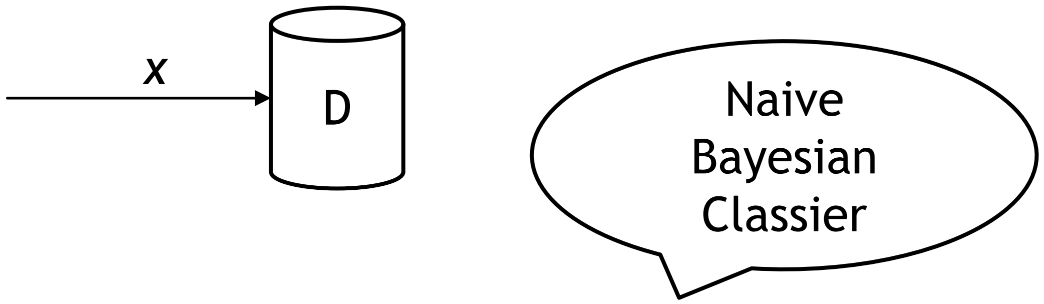
- We assume an EC scenario with computational resources spanning across different geographically distributed locations
- Local datasets $D_l, l = 1, 2, \dots, N$
  - $x =< x_1, x_2, \dots, x_M >$
- Detect the most significant features

Processing at the edge decrease latency

System Model (2)

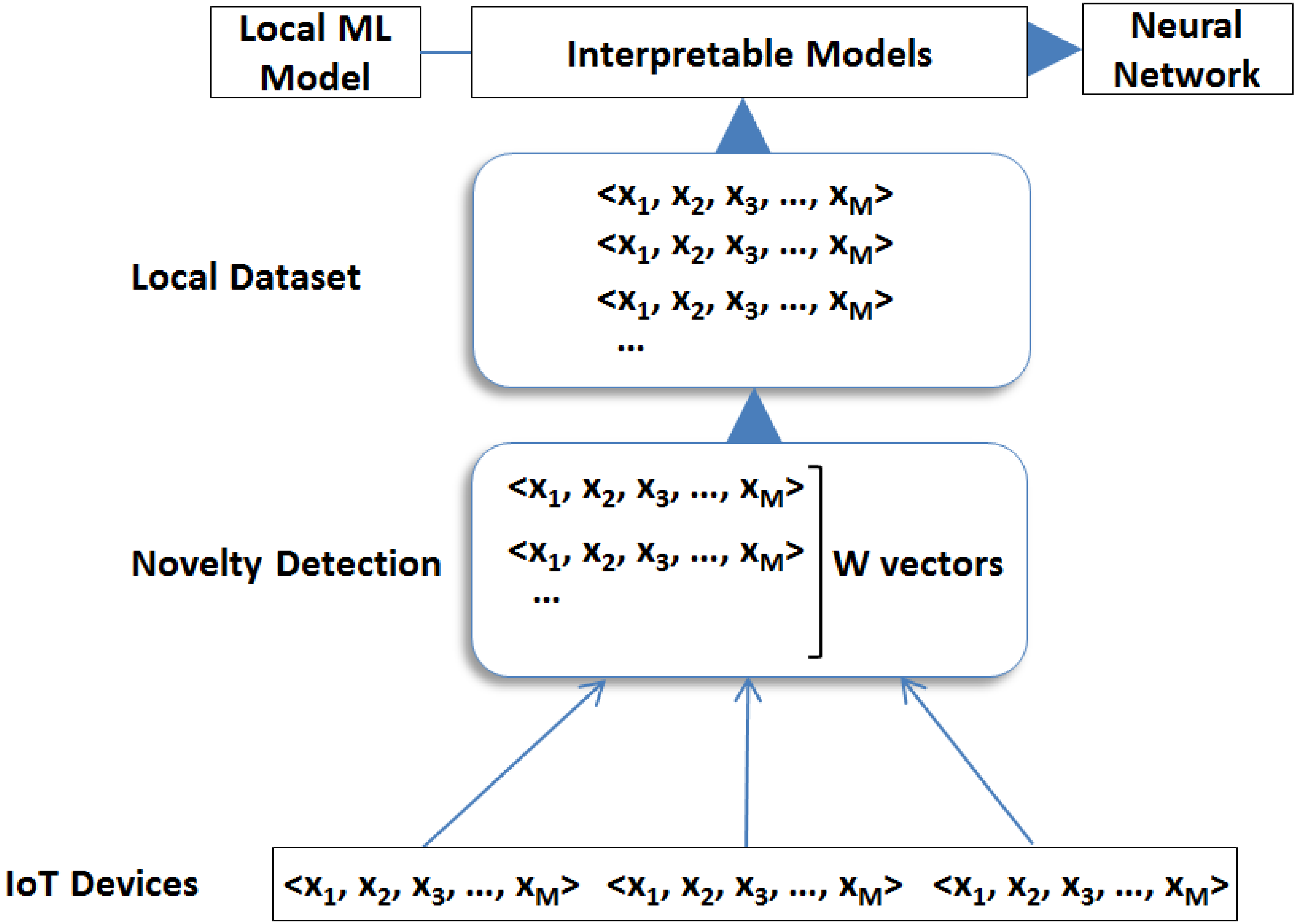
- Local datasets are characterized by specific statistical information
  - Mean
  - Standard Deviation

Accuracy is affected by the error between *D* and *x*



x deviates D → *reject x*

Architecture of an edge node



Performance Indicators

- Metric 1) Percentage of correct decisions Δ

$$\Delta = \frac{|\mathrm{CD}|}{|\mathrm{DS}|} * 100$$

$\Delta \rightarrow 100\%$ high accuracy

$\Delta \rightarrow 0\%$ low accuracy

CD

the set of correct decisions related to the storage of the appropriate data locally

DS

represents the set of decisions taken in our experimental evaluation

Performance Indicators

- Metric 1) Percentage of correct decisions Δ
- Metric 2) standard deviation of data σ

$$\sigma \downarrow \;\; \rightarrow \text{solid dataset}$$

$$\sigma \uparrow \;\; \rightarrow \text{unreliable dataset}$$

Performance Indicators

- Metric 1) Percentage of correct decisions Δ
- Metric 2) standard deviation of data σ
- Metric 3) average time τ required to take a decision

Performance Indicators

- Metric 1) Percentage of correct decisions Δ
- Metric 2) standard deviation of data σ
- Metric 3) average time τ required to take a decision

Experimental Setup

WS-DREAM datasets

- Response time
- Throughput
- 339 users
- 5,825 Web services.

$M \in \{20,50,100\}$
$W \in \{10\%,20\%,50\%\}$

Performance Assessment

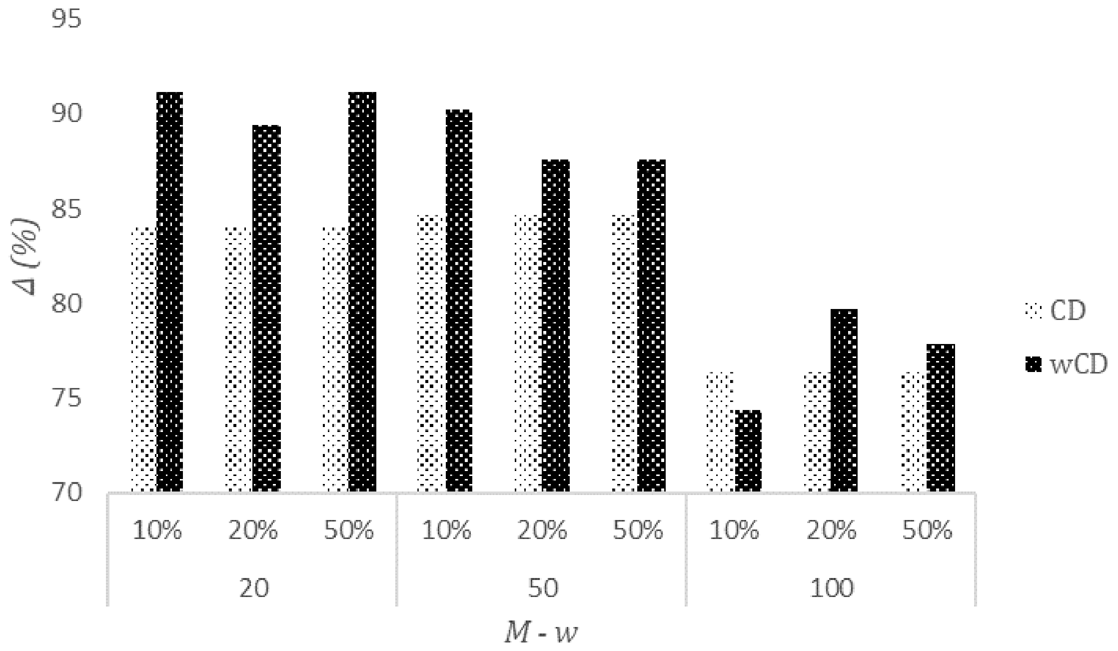


Fig. 1. Performance evaluation for the correct decisions derived by our model compared with the baseline solution (Naïve Bayes Classifier)

Performance Assessment

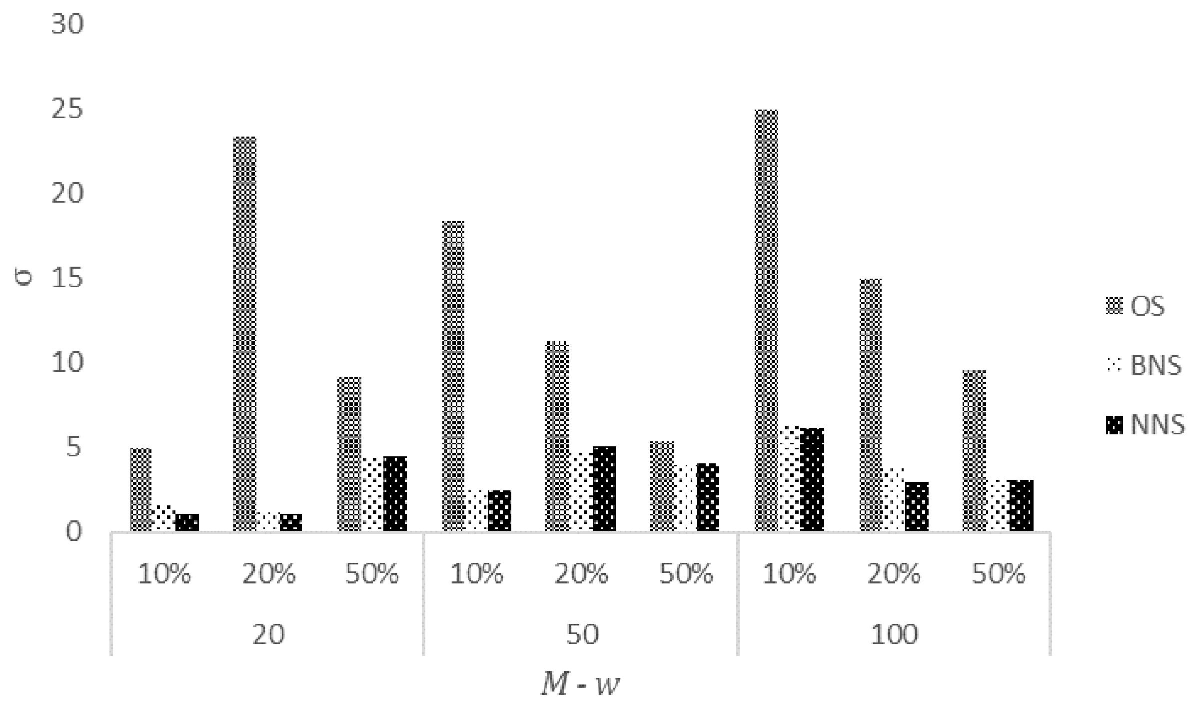


Fig. 2. Data solidity as delivered by the proposed model in comparison with other models found in the respective literature (i.e., the OS and the BNS models).

Performance Assessment

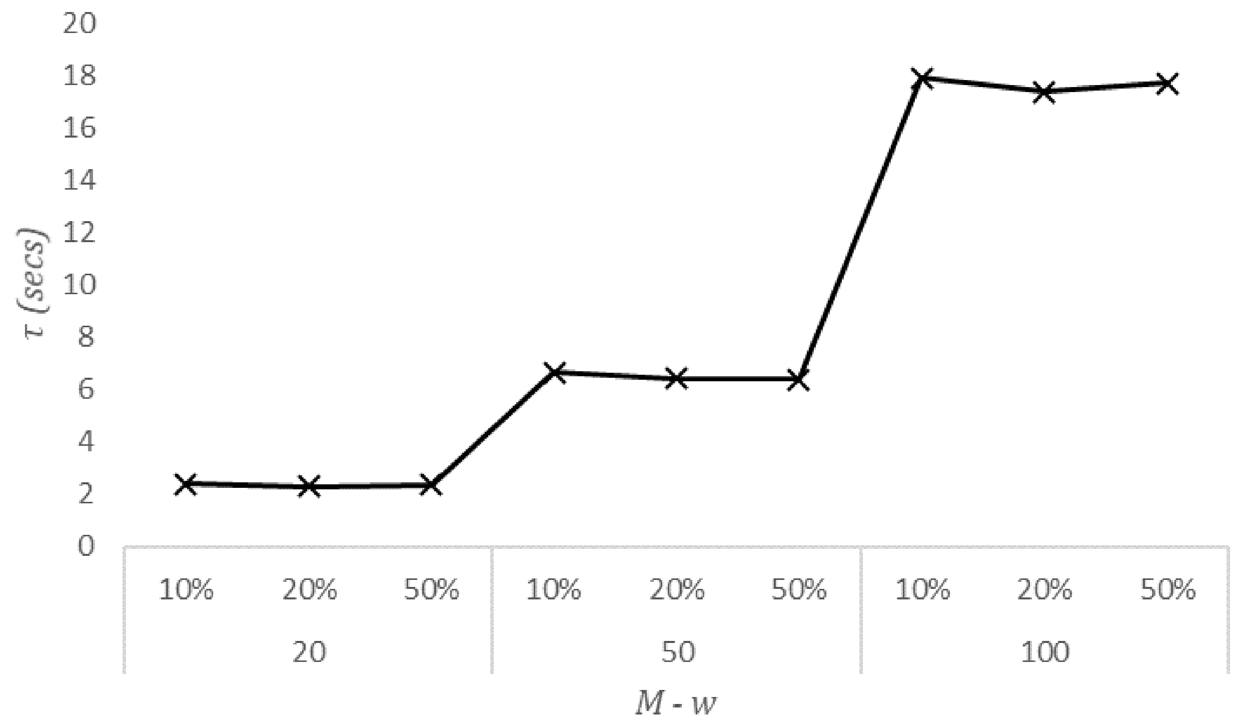


Fig. 3. Performance evaluation related to τ i.e., the time requirements for concluding the final sun-set of features

Performance Assessment

Table 1. Comparative results for the Δ metric.

M	w	wCD	PCA
	10%	91	71
20	20%	89	89
	50%	91	84
	10%	90	81
50	20%	87	83
	50%	87	84
	10%	83	78
100	20%	79	78
	50%	77	74

Future Work

- Cover the uncertainty around the significance of each feature
- Sliding window approach

Thank you!