Energy Efficient Data Collection via Supervised In-Network Classification of Sensor Data

Lorenzo A. Rossi, Bhaskar Krishnamachari and C.-C. Jay Kuo
Ming Hsieh Department of Electrical Engineering
University of Southern California, Los Angeles, CA 90089-2564, USA
E-mail: lrossi@usc.edu, bkrishna@usc.edu, cckuo@sipi.usc.edu

***Abstract*—In wireless sensor networks, data collection (or gathering) is the task of transmitting rounds of measurements of physical phenomena from the sensor nodes to a sink node. We study how to increase the efficiency of data collection via supervised in-network classification of rounds of measurements. We assume that the end users of the data are interested only in rounds characterized by certain patterns. Hence the wireless sensor network uses classification to select the rounds of measurements that are transmitted to the base station. The energy consumption is potentially reduced by avoiding the transmission of rounds of measurements that are not of interest to the end users. In-network classification requires distributed feature extraction and transmission. Such tasks can be less or more energy expensive than the transmission of measurements without classification. We provide analytical results and simulations on real data to show requirements and key trade-offs for the design of in-network data classification systems that can improve the collection efficiency. Besides, we study the impact of spatial subsampling of the sensor data (a way to further decrease energy consumption) on the classification performance.**

I. Introduction

Wireless sensor networks can be applied to the autonomous monitoring of ecosystems, industrial facilities, infrastructures (e.g. bridges or tunnels), cities, homes etc. We consider the scenario of a network of sensor nodes collecting measurements (*e.g.* temperature) from an ecosystem. The spatio-temporal recordings of the phenomenon provided by the sensor network can be seen as a sequence of images, where each pixel is associated to a different node.[1] The goal of this work is to study how in-network classification of the rounds of measurements can improve the energy efficiency of data collection. In this framework, the sensor network classifies each round of measurements to transmit only the rounds belonging to a class of interest to the end users. For example, a sensor network could be collecting daily rainfall data over a certain region and the end users could be interested only in rounds of data associated to very rainy days. This approach is potentially more energy efficient than transmitting each round of measurements to the base station, as long as the rounds of measurements of interest are infrequent and the classification task is not too energy demanding. The approach could also have a significant benefit in the scenario of a very large number of sensor networks (*e.g.* deployed in smart homes) transmitting data to a cloud server, since it could reduce the amount of data stored and processed at the server and therefore the fees due to the cloud provider.

The distributed extraction and classification of features from the sensing data require transmission of packets from the nodes and so consume energy. This constrains the number and the size of the feature packets that can be exchanged to classify the rounds of data. We show that the constraints depend on the network topology and the priors of the class of interest. We focus on the *intensity histogram* as a feature for the classification because its properties make the math of our analysis more treatable. We also study analytically and experimentally the impact of spatial subsampling of the sensor data on the classification performance, since it can decrease the cost to extract the features and transmit them to the sink.

There is a broad literature on supervised (and centralized) image classification [1]. Supervised and unsupervised data classification are also studied in the context of wireless sensor networks. The challenges inherent to distributed classification in wireless sensor networks are discussed in [2]. Alsheikh *et al.* in [3] present a broad overview of recent machine learning methods (including supervised classification) applied to wireless sensor networks. Another relevant topic in the field of learning (or classification) for wireless sensor networks is distributed *anomaly detection* [4], [5]. The main goals are to detect security attacks to a network and faulty measurements from the sensors. For many of the aforementioned works, it seems that the reduction of latency is the main motivation to perform classification at the sensor network, since it is not clear whether that is more energy efficient than just collecting measurements at the sink and then classifying them at the base station. To the best of our knowledge, there are no works that propose in-network data classification with the goal of efficiently selecting relevant rounds of sensor data to reduce the overall transmission to the end users.

Many works focus on compression methods (in particular on compressed sensing, *e.g.* [6], [7]) to increase the efficiency of data collection. Our approach can be integrated with compression methods for the transmission of measurements and features to the sink.

The color (intensity) histogram has been used in image processing for applications such as retrieval [1], medical imaging [8] and steganalysis [9]. The impact of image subsampling on

[1] Such pixels are not necessarily placed on a regular grid, due to the random placements of the nodes in the field.

histograms have been studied in [9] and [8], but with purposes and assumptions different from ours.

To the best of our knowledge, ours is the first attempt to study the problem of in-network classification for efficient data collection in sensor networks in a comprehensive manner. We claim the following contributions:

- we highlight key constraints and trade-offs to design in-network classification approaches for efficient data collection in wireless sensor networks
- we study the problem for a class of admissible features that satisfies energy efficiency constraints (the intensity histogram) and propose an algorithm for its distributed extraction
- we study the impact of spatial subsampling over in-network classification, provide analytical expressions for the related error metrics and validate them via simulations on real data.

The rest of this document is organized as follows. The problem formulation is given in Section II. Analytical conditions on the maximum allowed size for a class of admissible feature vectors are presented in Sec. III. An approach for the distributed feature extraction is given in Sec. IV. Sec. V studies analytically how spatial subsampling of the sensing frames can impact the performance of the classification via intensity histograms. Experimental results are provided in Sec. VI. Conclusive remarks are in Sec. VII.

II. Problem Formulation

Our goal is to design sensor network systems such that data collection driven by in-network classification is more energy efficient than data collection without classification. We define general constraints on the energy available for in-network feature extraction and classification.

A. System assumptions

We consider a set of N sensor nodes deployed over a compact region $\Omega \subseteq \mathbb{R}^q$, sampling a spatio-temporal process $Z(\mathbf{x}, t)$, $\mathbf{x} \in \Omega$, during a discrete temporal interval $t \in [t_0, ..., t_0 + k\Delta t, ...]$, $k = 1, 2, ..., K$. All the nodes are connected via a *data collection tree* to the sink. Packets containing samples of the field $Z(.)$ are transmitted to the sink via multi-hop communication through the data collection tree, without losses or collisions. The nodes transmit packets with fix communication range d. The sink transmits the measurements from the nodes to a remote base station. The nodes have limited energy supply, while the sink does not suffer of any relevant limitation in the energy supply[2]. We adopt the shortest path tree (SPT) as data collection scheme. Furthermore, we assume that the sampling time at the sensor nodes is synchronized [3] and that the network is faster than the measured phenomena. We also assume that for a node data transmission requires much more energy than the sensing and/or processing tasks.

Our analyses and simulations (Secc. V and VI) on the effect of data downsampling on the classification performance suggest that our proposed framework is robust to some degree to packet loss.

Definition 1: We say that a complete round of sensor measurements, $Z(\mathbf{x}_i, t)$, $i = 1, ..., N$ at time t is a *sensing frame*.

B. Problem formulation for error free classification

Without loss of generality, we assume that each sensing frame $Z\{(\mathbf{x_i}, t_0)\}$ acquired by the network at a certain time t_0 belongs to one of two different classes, S_1 and S_0, and that the end users are more interested in the sensing frames belonging to class S_1 (e.g. intense rain over Ω) than to those belonging to class S_0 (*e.g.* moderate or no rain). We also refer to S_1 and S_0 respectively as positive and negative class. Since we assume that the sink does not have energy limitations (*e.g.* because it is connected to the main electrical network), the only opportunity to save energy is in the multi-hop communication from the nodes to the sink. The network can save energy by limiting the number and/or the size of the messages sent from the nodes to the sink.

Let us denote with E_1 and E_0 the the energy spent by the network to transmit sensing frames, respectively of class S_1 and S_0, to the sink. With the assumption of shortest path data collection and fixed radio range communication, an expression to compute the energy cost E_1 is given in [11]:

$$E_1 = \sum_{n=1}^{N} r(n) d(\mathbf{x}_n), \tag{1}$$

where $r(n)$ and $d(\mathbf{x}_n)$ are respectively the size of the packet generated by the n-th node and the hop distance from that node to the sink. We set E_0 to be a fraction of E_1 according to a certain scalar factor $\alpha < 1$, i.e. $E_0 = \alpha E_1$. If $\alpha = 0$, the network does not transmit to the sink the samples classified as belonging to S_0.

We assume that over a temporal interval of observation $[1, K]$, there are K_1 and K_0 sensing frames of classes S_1 and S_0 respectively, with $K = K_1 + K_0$. The total energy spent by the network to perform data collection, assuming that there are no errors in the classification of the samples, is:

$$E = K_1 E_1 + K_0 E_0 + K E_a, \tag{2}$$

where E_a is the additional energy spent by the network at time K in order to classify a sensing frame. We assume that E_a depends only on the additional communication messages needed to analyze the data and not on the computation at the nodes. The additional communication packets contain for instance feature vectors, or data needed to extract features.

Let β be the ratio between the data classification cost and E_1, *i.e.* $\beta =: E_a/E_1$. To compare the cost of data collection via content classification and the the cost of regular data collection, we define the *efficiency factor*, η, as $\eta := E/(KE_1)$,

[2]The constraints on the energy available for classification would be more relaxed, if the sink node had limited energy supply.

[3]Network wide synchronization at the sub-millisecond level is possible at relatively small additional cost using protocols such as FTSP [10].

where E is given by equation (2). We can rewrite η as a function of α and β:

$$\eta = p_1 + \alpha(1 - p_1) + \beta, \tag{3}$$

where the term $p_1 := \frac{K_1}{K}$ is the frequency of occurrence of the sensing frames of high interest; p_1 can be seen as an estimate of the prior probability for the events of higher interest to the end users and it depends on the nature of the phenomenon monitored by the network. The factors α and β depend on the design of the system. Our goal is to design sensor network systems such that $\eta < 1$, *i.e.*: $p_1 + \alpha(1 - p_1) + \beta < 1$.

Definition 2: We say that an in-network classification system is **efficient** if $\eta < 1$.

We have implicitly assumed that the classification of a sensing frame is performed by the sink. When the sink classifies a sensing frame as positive, it sends requests to the nodes to transmit their measurements, We introduce a term, θ, to account for the cost of the request messages. We assume $\theta < 1$. For simplicity, we consider the case of $\alpha = 0$, *i.e*:

$$\eta = p_1(1 + \theta) + \beta \tag{4}$$

Proposition 1: A necessary condition for an in-network classification system to be efficient is that $\beta < 1$.

The proof of Prop. 1 is obvious. The relevant implication is that energy spent to transmit the features to the sink must be smaller than that energy spent to collect the sensing data. Some features whose extraction involves a lot of communication among the nodes may not fit the above constraint.

C. Formulation under misclassification errors

In a real situation, the system can classify some sensing frames incorrectly. We denote the false positive and false negative error rates respectively as ε_p and ε_n. Hence, to take into account the misclassification of sensing frames, the efficiency factor (4) can be written as:

$$\eta = p_1(1 + \theta) + \beta + [(1 - p_1)\varepsilon_n - p_1\varepsilon_p](1 + \theta) < 1 \tag{5}$$

The above expression provides the efficiency of in-network classification as a function of prior probabilities, error rates and costs of transmitting features and notifications. Intuitively, a high rate of false positives, ε_p, increases the amount of energy spent by the system, while a high rate of missed hits ε_n causes loss of information to the end users, although decreasing the energy spent by the network. There is a trade-off between the average cost factor for collecting the features, β, and the classification errors: intuitively, lowering β may increase the bias of the classifiers and hence the error rates.

III. Bound on the Feature Size

We make further assumptions on the distributed feature extraction task. We assume that each node aggregates the features extracted from its own measurements with the features received from its children nodes and then transmits the resulting feature vector to its parent node. Let v be the number of bits used to encode the feature vectors at each node. To simplify the math of our derivations, we assume v to be constant over the network. Figure 1 shows how in this case feature collection can be less expensive than measurement collection. Let p be the number of bits used to encode a standard measurement packet.[4] Let $\rho := v/p$. Based on the energy cost model we adopted in (1), the cost of sending a standard package to the sink is $p \times h \times d$, where h is the number of hops between the node and the sink and d is the maximum hop distance allowed by the fix radio range. The total cost of a full round of data collection for a fully connected sensor network is given by the expression: $E = \sum_{h=1}^{N_h} p \times n(h) \times h \times d$, where N_h is the number of hops and $n(h)$ is the number of nodes at the hop h. On the other hand, the transmission cost to classify the data is equivalent to the cost of transmitting the feature vectors hop by hop all the way to the sink, which can be expressed as: $E_a = \sum_{h=1}^{N_h} v \times n(h) \times d = vdN$, where N is the total number of nodes in the network. Recall that we have assumed v to be constant over the network. The necessary condition defined in Prop. 1 (Ssec. II-B) requires that $\beta = E_a/E < 1$. Therefore:

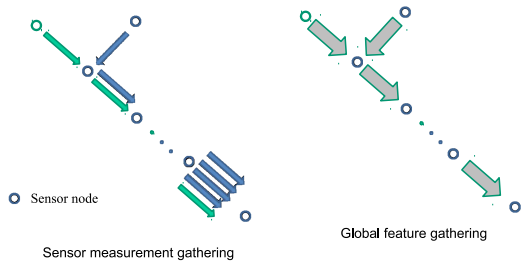


Fig. 1. Packet size for collection of sensor measurement vs. global feature collection. In the former case, the number/size of packets increases with the number of hops. In the latter, it remains constant.

$$\beta = \frac{vN}{p\sum_{h=1}^{N_h} n(h)h} = \rho\frac{N}{\sum_{h=1}^{N_h} n(h)h} < 1. \tag{6}$$

Throughout the rest of this work, we refer to β as the *feature to data collection cost ratio (FDR)*. By solving the eq. (6) for ρ, we have :

$$\rho < \frac{\sum_{h=1}^{N_h} n(h)h}{N}. \tag{7}$$

Hence, **the maximum allowed size for a feature vector depends on the number of hops and on the spatial distribution of the nodes**. If due to the problem constraints, we need to chose ρ so that $\beta \leq \beta_0 < 1$, we have:

$$\rho \leq \beta_0 \frac{\sum_{h=1}^{N_h} n(h)h}{N}.$$

Example 1: Consider the case of a sensor network whose topology consists of a balanced binary tree, *i.e.* each node has

[4]Note that both v and r include also a header messages.

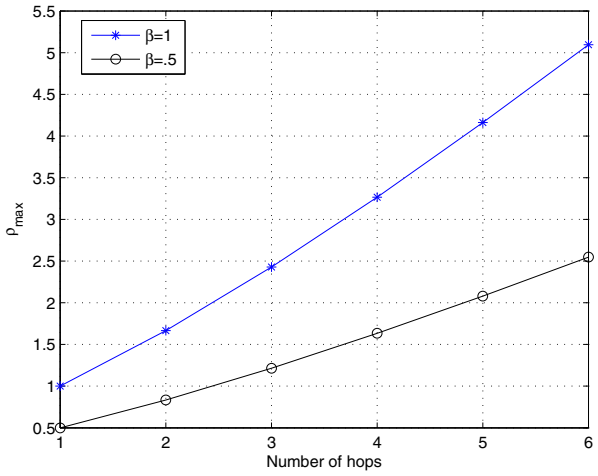


Fig. 2. Maximum allowed number of features vs. number of hops for an efficient classifier in a binary balanced tree network.

exactly 2 children and $n(h) = 2^{h-1}$. From eq. (6), we have:

$$\rho < \frac{\sum_{h=1}^{N_h} 2^{h-1} h}{\sum_{h=1}^{N_h} 2^{h-1}} = \frac{2^{N_h}(N_h - 1) + 1}{2^{N_h} - 1}$$

In Figure 2, we show the value of ρ as a function of the number of hops for $\beta = 1$ and $\beta = .5$ respectively. For instance for a network with 4 hops, and $\beta = .5$, $\rho = 1.6$.

Therefore, in-network content classification may not be convenient for sensor networks with a small number of hops (e.g. 2). Even in networks with 4 or more hops, the size of the feature vectors is strictly limited, thus with risk of bias in the classification. This also poses restrictions on the kind of features allowed.

IV. Classification via Intensity Histograms

In Sec, III, we assume that the size of the feature vector is constant throughout the extraction process over the network. A type of feature that respects this constraints is the *intensity histogram*, which has been widely used in image retrieval [1]. The histogram of two distinct subsets of measurements is equal to the sum of the histograms of the two subsets. Therefore, we focus the rest of this work on the classification of sensing frames via intensity histograms.

We give some formal definitions. Let $\mathbf{Z} \in \mathbb{N}^N$ a vector of sensing measurements, with $N \in \mathbb{N}$ and $z(i) \in (a, b) \subset \mathbb{N}$, where $i = 1, 2 \ldots, N$ and (a, b) is a finite subinterval[5]. The elements of $\mathbf{Z}$ are the measurements extracted at a time t_k by a set of N sensing nodes deployed over a compact region Ω.

Definition 3: We define the *histogram* of $\mathbf{Z}$ as $\mathbf{h}_Z$: $(a, b) \rightarrow \mathbb{N}^N$ *s.t.*:

$$h_Z(u) := \sum_{i=1}^{N} x_i, \tag{8}$$

[5]Sensing measurements are quantized and can be assumed integer up to a constant scale factor

$\forall u \in (a, b)$, where

$$x_i = \begin{cases} 1 & \text{if } z(i) = u, \\ 0 & \text{if } z(i) \neq u. \end{cases} \tag{9}$$

Usually the number of bins is smaller than the number of possible values of z.

Definition 4: We define the *$L-$bin histogram* of $\mathbf{Z}$ as $\tilde{\mathbf{h}}_Z$: $(a, b) \rightarrow \mathbb{N}^L$ *s.t.*: $\tilde{h}_Z(k) := \sum_{i=1}^{N} \tilde{x}_i$, $\forall k \in (1, L)$, where $\tilde{x}_i = 1$ if $z(i) \in I_k$, or $\tilde{x}_i = 0$ if $z(i) \notin I_k$, where the subintervals I_k, $k = 1, \ldots, L$ form a partition of the finite interval (a, b), *i.e.*: $I_k \subset (a, b)$, $\cup_k I_k = (a, b)$ and $I_k \cap I_l = \varnothing$, $k \neq l$.

The intensity histograms of a sensing frame can be extracted in a distributed manner from the sensor nodes. The sink classifies a sensing frame by comparing the distance of its histogram with a set of *reference histograms*, *i.e.* via a linear classifier. Each reference histogram represents a class of sensing frames. The reference histograms can be computed offline at the base station from a set of historical data.

V. Effect of Subsampling on Classification

A way to decrease the feature acquisition cost, and to potentially increase the efficiency of the system, is to reduce the number of messages to analyze the data, and so the number of bits per bin, by subsampling the sensor measurements used to compute the histogram. We analyze how the histograms of the frames are affected by subsampling od the rounds of measurements (Ssec. V-A) and how this impacts the classification performance (Ssec. V-B).

A. Impact of subsampling over intensity histogram

We model the histogram of a subsampled sensing frame as a random vector, whose statistics are characterized by the histogram of the sensing frame before subsampling.

Let $\mathbf{Z}$ be a N-dimensional vector of sensing measurements. See Sec. IV (Deff. 3 and 4) for the formal definition of the histogram of $\mathbf{Z}$. Let $\mathbf{Y} \in \mathbb{N}^m$, $m < N$ a subsampled version of $\mathbf{Z}$ according to a subsampling factor D. In the analysis and the experiments, we consider three types of **subsampling**:

Definition 5: Regular subsampling: $y(i) = z(Di)$, where D is the subsampling factor; $m = \lfloor \frac{N}{D} \rfloor$.

Definition 6: Uniform random subsampling: $\mathbf{Z}$ is sampled at m random locations, uniformly distributed over the N locations of $\mathbf{Z}$; $m = \lfloor \frac{N}{D} \rfloor$.

Definition 7: Bernoulli subsampling: each element $z(i) \in \mathbf{Z}$ can be sampled with probability $p = \frac{1}{D}$. In this case, $E[m] = \lfloor \frac{N}{D} \rfloor$. This methodology is easy to implement in a distributed system and therefore suitable for wireless sensor networks.

We also define a subsampling operator.

Definition 8: Let $\mathbf{Z}$ be an N dimensional set of measurements, D an integer s.t. $N/D = m$ is also an integer. We define a *subsampling operator* $\mathcal{S}_D(.)$ s.t. $\mathcal{S}_D(Z) = \mathbf{Y}$, where Y is a subsampled version of Z with subsampling factor D.

Let the $m-$dimensional vector $\mathbf{Y} = \mathcal{S}_D(Z) =$ be a subsampled version of $\mathbf{Z}$ with subsampling factor D, $(m < N)$. Let $\mathbf{h}_Z$ and $\mathbf{h}_Y$ be respectively the histograms of $\mathbf{Z}$ and $\mathbf{Y}$. Our

goal is to analyze the relationship between $\mathbf{h}_Y$ and $\mathbf{h}_Z$, where $\mathbf{Z}$ is the sensing frame prior subsampling. The histogram, $\mathbf{h}_Y$ of the subsampled process $\mathbf{Y}$ can be seen as the realization of a random vector $\mathbf{H}_Y$. We formulate the following proposition.

Proposition 2: We assume for simplicity that $m, N \in \mathbb{N}$ *s.t.* $\frac{m}{N} \in \mathbb{N}$. Let $\mathbf{Y}$ be a uniform random subsampled version of $\mathbf{Z}$ and $\mathbf{h}_Y$ the histogram of $\mathbf{Z}$. Then:

$$E[\mathbf{H}_Y] = \frac{1}{D}\mathbf{h}_Z, \tag{10}$$

where $E[.]$ is the statistical expectation.

Proof: We evaluate the element $\mathcal{H}_Y(u)$ of the random vector $\mathbf{H}_Y$. From the definition of histogram and the hypotheses:

$$\begin{aligned} E[\mathcal{H}_Y(u)] &= \sum_{i=1}^{m} E[x_i] = mP\left(Y = u\right) \\ &= \frac{N}{D}\frac{1}{N}h_Z(u) = \frac{1}{D}h_Z(u), \end{aligned}$$

where x_i is defined in Eq. (9), $P(Y = u) = \frac{1}{N}h_Z(u)$ thanks to the hypothesis of uniform random sampling. Therefore: $E[\mathbf{H}_Y] = \frac{1}{D}\mathbf{h}_Z$. ■

Proposition 3: Equation (10) applies also to the case of Bernoulli subsampling.

Proof: We consider an element of the random vector $\mathcal{H}_Y(u_0)$, *i.e.*:

$$E[\mathcal{H}_Y(u)] = E\left[\sum_{i=1}^{N} x_i\right] = \frac{N}{D}P(Y = u) = \frac{1}{D}h_Z(u),$$

$$x_i = \begin{cases} 1 & \text{if picked with } P = \frac{1}{D} \\ 0 & \text{if not picked with } P = 1 - \frac{1}{D} \end{cases}$$

■

Corollary 1: Let $\mathbf{Z}$ be an $N-$dimensional set of measurements with histogram $\mathbf{h}_Z$ and $D \in \mathbb{N}$, s.t. $m = N/D \in \mathbb{N}$, $\mathbf{Y} = \mathcal{S}_D(\mathbf{Z})$. Then

$$\text{as } N \to \infty,\ \mathbf{h}_Y \to \frac{1}{D}\mathbf{h}_Z \text{ with probability } P = 1. \tag{11}$$

Proof: Let $\{\mathbf{Z}_i\}$, $i = 1, 2, ...,$ be a sequence of N_i-dimensional vectors s.t. $\{\mathbf{h}_{Z_i}\} = \{\frac{N_i}{N_0}\mathbf{h}_{Z_0}\}$. Let $\{\mathbf{Y}_i\}$ a sequence of m_i dimensional vectors such that $\{\mathbf{Y}_i\} = \{\mathcal{S}_D(\mathbf{Z}_i)\}$. We evaluate an element of $E[\mathbf{H}_{Y_i}]$. From Proposition 2:

$$E[h_{Y_i}] = \frac{1}{D}h_{Z_i} = \frac{N_i}{D}\frac{h_{Z_0}}{N_0} \tag{12}$$

By dividing the left and right hand sides with N_i, we see that all the elements of the random sequence $\{h_{Y_i}/N_i\}$ have the same expected value, $E[\frac{h_{Z_0}}{DN_0}]$. Due to the strong law of large numbers, we have that: $h_{Y_i} \to \frac{N_i}{DN_0}h_{Z_0}$, as $i \to \infty$. Now let $\mathbf{h}_Z := \frac{N_i}{N_0}\mathbf{h}_{Z_0}$ and $N = N_i$. Then $i \to \infty$ implies $N \to \infty$, hence the proof. ■

We can see $\mathbf{h}_Y$, the histogram of a subsampled version of $\mathbf{Z}$, as the estimate of a scaled version of $\mathbf{h}_Z$. A small number of available samples N implies the likelihood of a larger estimation error. This estimation error impacts the classification of the subsampled sensing frames.

Proposition 4: Propositions 2, 3 and Corollary 1 hold also for the case of $L-$bin histograms.

Proof: Let $\mathbf{h}_Y$ be the histogram of $\mathbf{Y}$ and $\tilde{\mathbf{h}}_Y$ the $L-$bin histogram $\mathbf{Y}$. We observe that for a certain interval $I_j = (a_j, b_j)$: $\tilde{h}_Y(j) = \sum_{i=a_j}^{b_j} h_Y(i)$. Hence, the proofs follow by observing that: $E[\tilde{h}_Y(j)] = \sum_{i=a_j}^{b_j} E[h_Y(i)]$. ■

From evidence gathered via simulations (Sec. VI), we make the following conjectures.

Conjecture 1: Let $\mathcal{E}^2$ be the mean square error in the estimate of the scaled histogram of $\mathbf{Z}$. If $D < 1$,

$$\mathcal{E}^2 = \left\| \frac{\mathbf{h}_Z - D\mathbf{h}_Y}{N} \right\|^2 \tag{13}$$

Then

$$E[\mathcal{E}^2] = O\left(\frac{1}{m}\right). \tag{14}$$

Recall that $m = N/D$ is the number of samples for $\mathbf{Y}$.
If the number of nodes deployed is relatively small (*e.g* $N < 50$), we are interested in predicting the MSE for m small. Therefore:

Conjecture 2: An approximation of the expected value of the MSE, Eq. (13), is given by:

$$E[\mathcal{E}^2] \approx \frac{1}{m}\left(1 - \frac{1}{D}\right)^2. \tag{15}$$

If m is small (*i.e.* the subsampling factor D is large), Eq. (15) is affected by the term $1/m$. $\left(1 - \frac{1}{D}\right)^2$ is zero when $D = 1$ (Fig. 6).

B. Analytical computation of the error metrics

We propose analytical expressions to predict the error rates for histogram based classification of sets of labeled sensing frames subsampled via uniform random or Bernoulli subsampling. Let $\{\mathbf{Z}_i\}$, $i = 1, 2, \ldots, N_t$, a sequence of N dimensional sensing frames, $\{\mathbf{Y}_i\}$, $i = 1, 2, \ldots, N_t$, the corresponding sequence of m_i dimensional[6] subsampled sensing frames, with $\{\mathbf{h}_{Z_i}\}$ and $\{\mathbf{h}_{Y_i}\}$ the corresponding sequences of L dimensional histograms. We can see all the histograms as laying on hyperplanes in an L dimensional space, since $\sum_{j=1}^{L} h_{Z_i}(j) = N$ and $\sum_{j=1}^{L} h_{Y_i}(j) = m_i$.

We assume that each element of the sequence of histograms associated to $\mathbf{Z}_i$ (and hence the sensing frames $\mathbf{Z}_i$ themselves) either belongs to a class S, or to its complement, $\bar{S}$. The process of subsampling a sensing frame $\mathbf{Z}$ and then to extract the intensity histogram can be modeled as the mapping of the histogram $\mathbf{Z}$ onto a different hyperplane in an L dimensional space. In the case of uniform random or Bernoulli subsampling of a frame $\mathbf{Z}$, there is a finite set of possible subsampled versions, $\mathbf{Y} = \mathcal{S}_D(\mathbf{Z})$, and so of possible histograms $\mathbf{h}_Y$. Some of these histograms may belong to class S, while the others to class $\bar{S}$. We indicate with $\{\mathbf{h}_{Y_i}^{(l)}\}$, $l = 1, \ldots$ the sequence of all the possible histograms that can obtained by subsampling $\mathbf{Z}_i$, where $\mathbf{Y}_i = \mathcal{S}_D(\mathbf{Z}_i)$. Let $S \ni \mathbf{Z}_i$.

[6] $m_i = m\ \forall i$ for regular and uniform random subsampling.

In the case of uniform random subsampling and $2-$bin histograms, the total error probability can be written as:

$$P_U(e) = \frac{1}{N_t \binom{N}{m}} \sum_{i=1}^{N_t} \sum_{l:\mathbf{h}_{Y_i}^{(l)} \in \bar{S}, \mathbf{h}_{Z_i} \in S} \binom{h_{Z_i}(1)}{h_{Y_i}^{(l)}(1)} \binom{N - h_{Z_i}(1)}{m - h_{Y_i}^{(l)}(1)}, \quad (16)$$

where recall that $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ and $m = \lfloor \frac{N}{D} \rfloor$.

For Bernoulli sampling with subsampling factor D and $2-$bin histograms, the total error probability is given by:

$$P_B(e) = \frac{1}{N_t} \sum_{i=1}^{N_t} \sum_{m=0}^{N} \sum_{l:\mathbf{h}_{Y_i}^{(l)} \in \bar{S}, \mathbf{h}_{Z_i} \in S} \binom{h_{Z_i}(1)}{h_{Y_i}^{(l)}(1)} \binom{N - h_{Z_i}(1)}{m - h_{Y_i}^{(l)}(1)} \left(\frac{1}{D}\right)^m \left(1 - \frac{1}{D}\right)^{N-m} \quad (17)$$

The expressions (17) and (16) require the *a priori* knowledge of the histograms of the sequence of sensing frames before subsampling, $\{Z_i\}$ and can be easily generalized to the case of multiple bin histograms. The (17) and (16) can be also used to compute error probabilities for missed hits and false positives (if they are applied respectively only to histograms associated to the class of interest or to its complement). Figure 7 shows a comparison between the analytical and experimental error rates.

If we have a certain population of N_1 histogram prototypes $\mathbf{H}_{Z_i}$, each of them with probability of occurrence $P(\mathbf{H}_{Z_i})$, we can rewrite the total error probability expression (18) as:

$$P_U(e) = \binom{N}{m}^{-1} \sum_{i=1}^{N_h} P(\mathbf{H}_{Z_i}) \sum_{l:\mathbf{H}_{Y_i}^{(l)} \in \bar{S}, \mathbf{H}_{Z_i} \in S} \binom{h_{Z_i}(1)}{h_{Y_i}^{(l)}(1)} \binom{N - h_{Z_i}(1)}{m - h_{Y_i}^{(l)}(1)} \quad (18)$$

A similar generalization can be made also for equation (16).

We can similarly compute analytical expressions for the average mean square error of a certain set of frames given their histograms. It can be shown that for two bin histograms and *uniform random* subsampling, the expected value of the MSE is given by:

$$E\left[\left\|\frac{\mathbf{h}_Z - D\mathbf{h}_Y}{N}\right\|^2\right] = \frac{1}{N_t}\binom{Dm}{m}^{-1} \sum_{i=1}^{N_t} \sum_{k=0}^{min(m, h_{Z_i}(1))} \binom{h_Z(1)}{k} \binom{Dm - h_Z(1)}{m-k} \frac{(Dk-m)^2}{D^2 m^2}, \quad (19)$$

where we assume that $N = Dm$ and that $h_{Z_i}(2) \geq h_{Z_i}(1)$. The (19) can be easily modified for the case $h_{Z_i}(2) < h_{Z_i}(1)$.

In the case of of Bernoulli Random subsampling, we have:

$$E\left[\left\|\frac{\mathbf{h}_Z - D\mathbf{h}_Y}{N}\right\|^2\right] = \frac{1}{N_t} \sum_{i=1}^{N_t} \sum_{m=0}^{\frac{N}{D}} \binom{N}{m}^{-1} \sum_{k=0}^{min(m, h_{Z_i}(1))} \binom{h_Z(1)}{k} \binom{N - h_Z(1)}{m-k} \frac{(Dk-m)^2}{D^2 m^2} \left(\frac{1}{D}\right)^m \left(1 - \frac{1}{D}\right)^{N-m}. \quad (20)$$

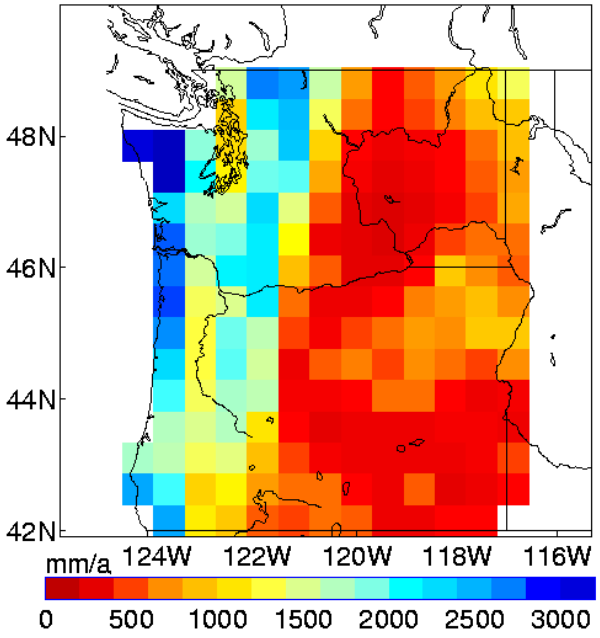


Fig. 3. The annual precipitation (mm, 1961-90) map of the Pacific Northwest of the United States. The coast is subject to much higher precipitation than the rest of region. Plot by Widmann *et al.* JISAO/U. of Washington [12].

VI. Experimental Results

For the experiments, we use a **rainfall dataset** provided by the University of Washington [12]: it gives the daily concentration of rainfall, between 1949 and 1994, for multiple observation points in the Pacific Northwest of the U.S. (Washington and Northern Oregon). The observation points are remapped onto a regular grid with 50Km distance between two contiguous points. We generally consider periods of observation between 4 and 7 years to avoid eventual trends affecting longer periods of time. Figure 3 shows the annual precipitation over the region.

We use **unsupervised learning**, via $k-$means clustering, to generate **labels** for training and testing sensing frames. We generally assume 2 classes and pick the sensing frames of the smallest cluster as representatives of the **class of interest**. Intuitively, we expect $k-$means to discriminate between sensing frames associated to rainy days and those associated to little or no rain. We decide a priori the number of bins for the histograms (*i.e.* the number of features). A subset of the labeled sensing frames is used to **train** a simple linear classifier. Unsupervised learning is performed one more time over the training subset. The centroids of the clusters (which are histograms) are used as reference features by the sensor network for the classification task in the testing stage. The **testing** stage is performed over the remaining portion of sensing frames. We expect that in an application scenario the *training* of the classifiers is performed offline, at the base station, with previously collected sets of measurements and the resulting parameters of the classifier are then transmitted to the sensor network. It seems reasonable to collect and study

the sensor data for a certain period of time (possibly weeks), before relying only the automatic classification of the data.

Note that both labeling (via k-means) and classification methods are based on linear partition of the feature space. Hence we expect a relatively low error rate in the testing stage. Our goal is to understand key trade-offs for this approach, not to claim excellence in classification performances.

Experiment 1: We use synthetic data to have full control over the spectral properties of the signals being sampled. We consider a vector $\mathbf{X}$ of N random samples drawn from zero a mean, unitary variance Gaussian distribution, $N(0,1)$. We process $\mathbf{X}$ via a 30 dimensional finite impulse response (FIR) filter (with symmetric spectrum) to obtain low, high and all-pass colored sequences $\mathbf{Z}$. The filter coefficients are computed via the Parks-McClellan algorithm.

We study the behavior of the mean square error metric as defined in (13):

$$E[\mathcal{E}^2] = E[\frac{1}{N^2}\,\|(\mathbf{h}_Z - D\mathbf{h}_Y)\|^2].$$

We average the results over 500 Monte Carlo trials for the case of regular subsampling, and 10000 trials for uniform random and Bernoulli subsampling. We consider 10 bin histograms as we vary number of samples N and subsampling rate D.

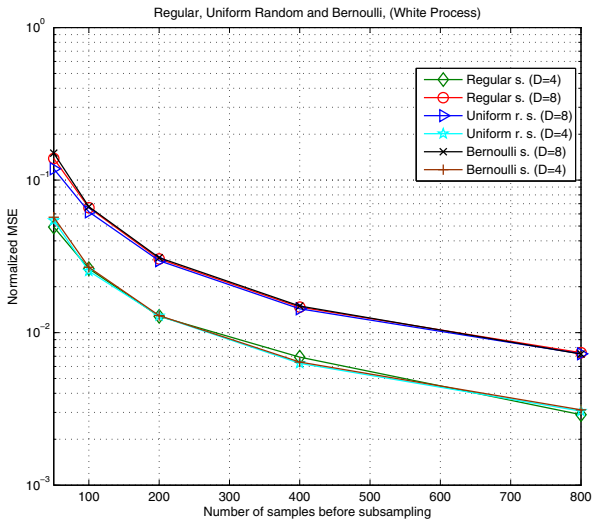


Fig. 5. Comparison of MSE for different types of regular, uniform and Bernoulli sampling for white processes.

Figures 4.a and 4.b compare the average MSE respectively for regular and uniform random subsampling of white, low pass and high pass processes. The average MSE for regular, uniform random and Bernoulli subsampling of a white process is shown in Fig. 5. Finally, Figure 6 compares the formula to approximate the average MSE (15). We observe the following:

- Aliasing has a slight effect on the MSE for regular sampling (Fig. 4.a), but no impact on the MSE for uniform random subsampling (Fig. 4.b). This seems due to the spatial sparsity of the sensed signals and the fact the random sampling is more effective than regular sampling because its a form of incoherent sampling [13].
- The values of the average MSE for Bernoulli and uniform random subsampling are very close (Fig. 5).
- The MSE decreases as $\frac{1}{m}$, where m is the number of samples of a subsampled sensing frame (Fig. 6).

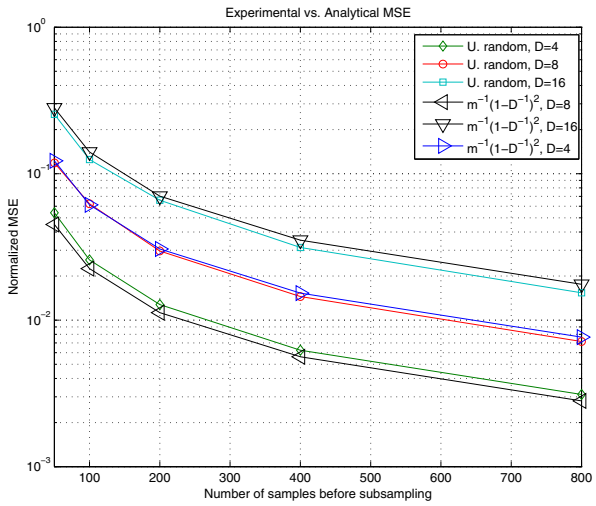
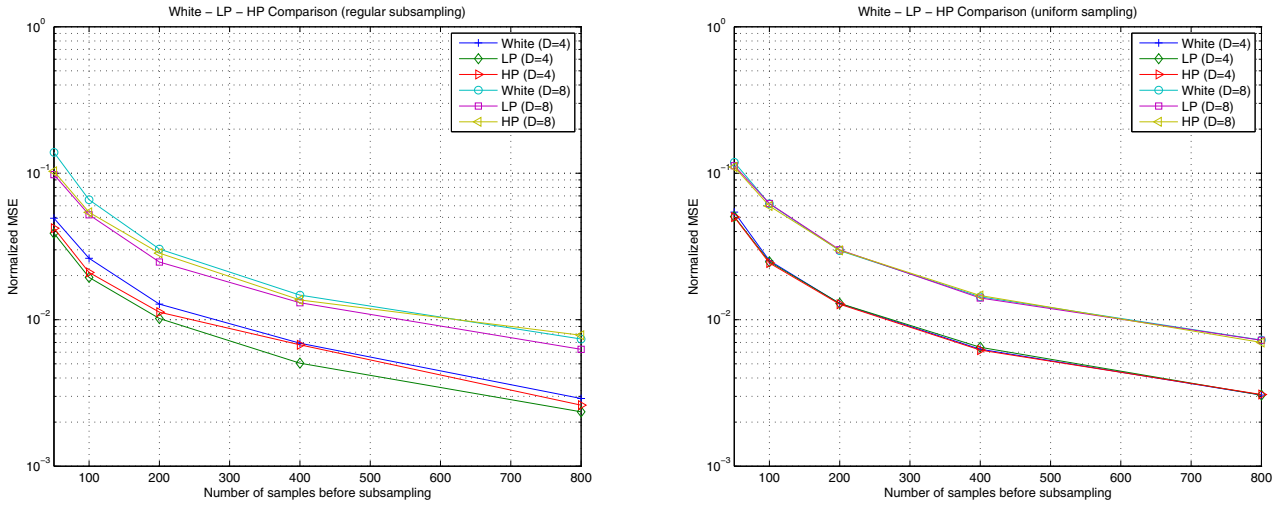


Fig. 6. Comparison of experimental MSE vs. $1/m$. approximation

The number of available samples m seems to be the parameter more influential on the normalized mean square error.

Experiment 2: We compare the error rates for regular, Bernoulli and uniform random subsampling over a 5 years period of the rainfall dataset. We consider 2 classes and 2 bin histograms. In Figures 7.a and 7.b we compare the error rates obtained analytically (16, 17) and experimentally respectively for uniform random and Bernoulli subsampling. Figure 8 shows the comparison of the analytical (19, 20), approximated (15) and experimental MSE. The analytical MSE matches the experimental one very accurately, while there is a small bias w.r.t. the approximated MSE. Figure 9 shows the error rates for the three types of subsampling on the same subset of the rain fall set, over a 5 year period. The experiments were repeated 500 times for uniform and Bernoulli subsampling.

In this set of experiments uniform random sampling leads to lower error rates than Bernoulli sampling. We believe that this is due to the fact that the number of samples in Bernoulli sampling is not constant over the trials and hence the trials whose number of samples is low contribute to increase the average error rates. Figure 9 shows also that regular subsampling presents lower missed rates than the other two methods, when the subsampling factor is 4. We believe that this has to do with the non uniform spatial distribution of the precipitation (see Fig. 3): *i.e.* the coast is much more rainy the rest of the region. The false positives here are actual dry days that are classified as rainy. Some random sampling trials may capture a relatively large number of rainy points on overall dry days and therefore get misclassified as rainy (false positives),



(a) (b)

Fig. 4. The average MSE for regular (a) and uniform random (b) subsampling vs. number of samples and signals with different spectral characteristics.

while regular subsampling with rate $1/4$ is more robust to this type of errors.

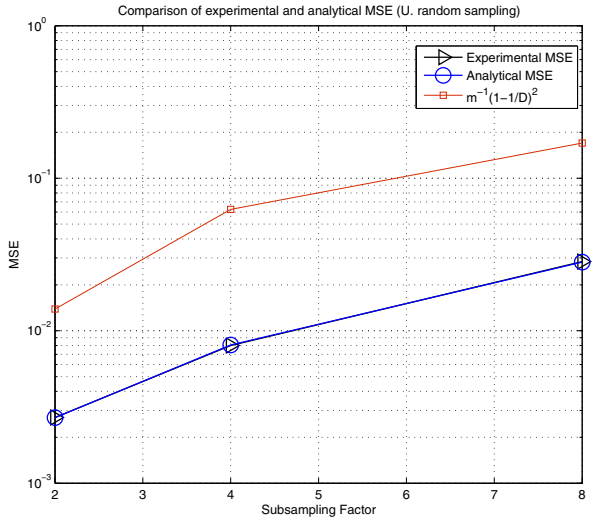


Fig. 8. Comparison of analytical, approximated and experimental average MSE for uniform random sampling over a 4 years period of the rainfall data set (2 bin histograms).

Experiment 3: The goal of this experiment is to estimate the impact of subsampling over the efficiency factor, η (3). Recall that we want η to be < 1 and it is given by:

$$\eta = p_1(1+\theta) + \beta + [(1-p_1)\varepsilon_n - p_1\varepsilon_p](1+\theta) < 1.$$

In this context, the reason to subsample is to increase η by decreasing the size of the histogram packets and therefore decreasing the feature to data collection cost ratio (FDR, *i.e.* β).[7] We use η and the missed hits rate, ε_M, to display the results of our experiments.

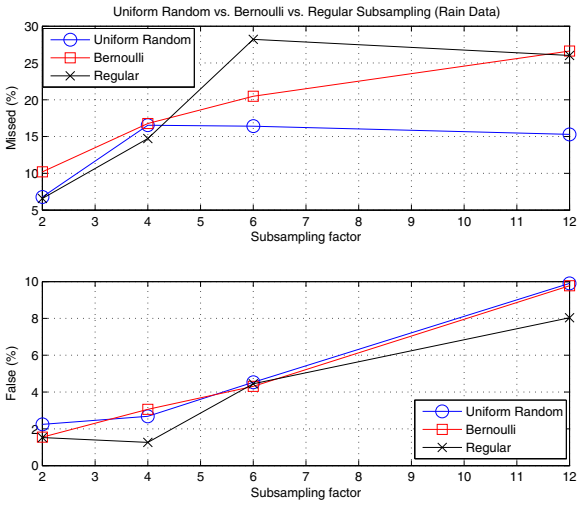
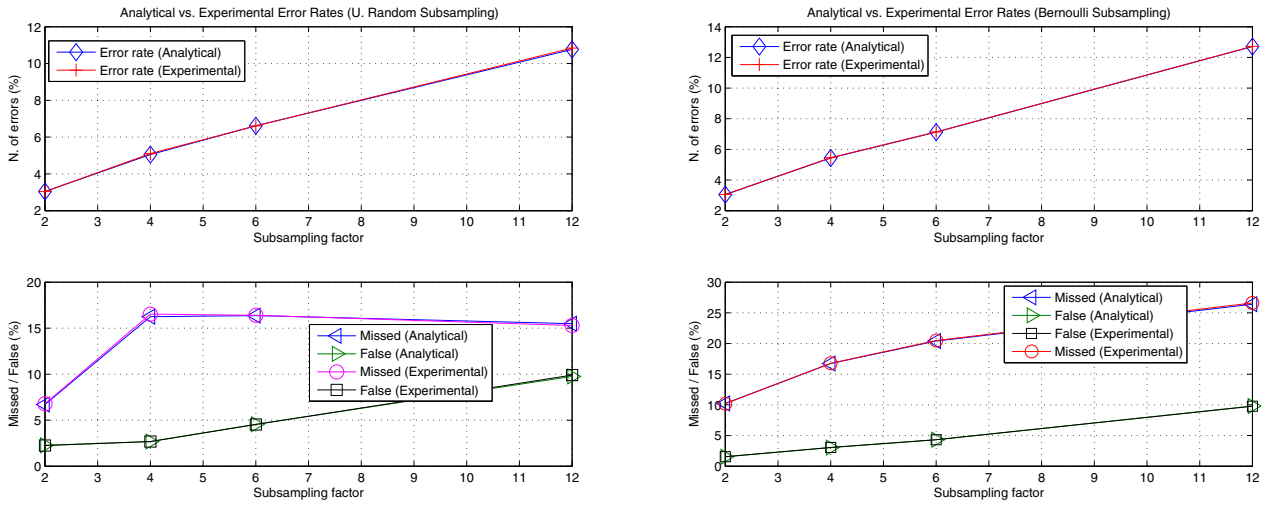


Fig. 9. Comparison of false negative (top) and false positive (bottom) rates for regular, Bernoulli and uniform random subsampling (rainfall data set, 5 years period, 2 bin histograms).

We consider 1 year of training frames and 3 years of testing frames over the whole region (163 nodes). The topology of a 6 hop network is shown in Figure 12. Recall that the nodes in the original dataset are placed over regular grid. A 16 bit header is assumed for all the packets. First, $8-$bin histograms are used for classification. Then we perform labeling, training and testing for the same sequences of sensing frames via 4 bin

[7]Note that subsampling reduces also the number of messages transmitted, when the measurements from the leaf nodes (*i.e.* without children) are not used. In this case, the nodes don't generate any packet.



(a) (b)

Fig. 7. Comparison of analytical and experimental error rates for uniform random (a) and Bernoulli (b) subsampling over a 5 years period of the rainfall data set (2 bin histograms). The false positive and false negative (missed hits) are at the bottom of each plot; the overall errors are at the top.

histograms. The nodes performs Bernoulli subsampling of the sensing signals: at each data acquisition time t_k, each node uses its data for the distributed computation of the histogram with probability $1/D$, where D is the subsampling factor. Bernoulli subsampling is easily implementable in a distributed way in any type of sensor network architecture, since each node needs only to know D. We consider also regular and uniform subsampling to better understand the behavior of the system.

Figure 10 compares the error rates and the effective efficiency for regular, uniform random and Bernoulli subsampling. Regular subsampling leads to slightly better performances than the other two methods. Figure 11 displays η for Bernoulli subsampling, for various header sizes and subsampling factors. It shows that the effective efficiency improves as the header size increases from 0 to 128 bits. On the other hand, the sensitivity of η to the subsampling factor decreases as the header size increases. Therefore, networks whose protocols require packets with large headers can perform efficient in-network classification without the need large subsampling factors. Subsampling becomes necessary with protocols requiring packets small headers (8 bits or less in this example).

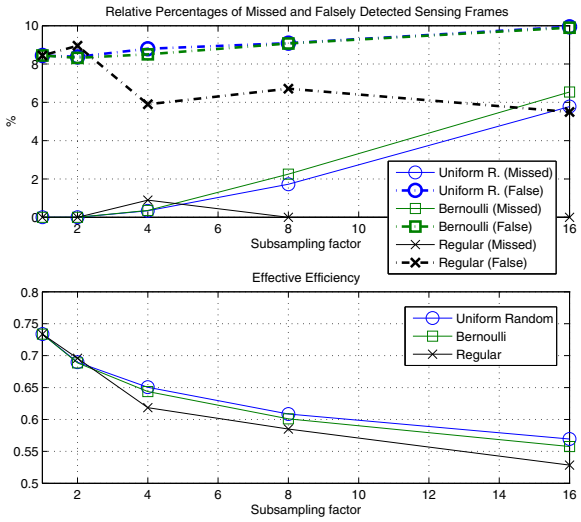


Fig. 10. Error rates for sensing frame classification via $4-$bin histograms.

VII. Conclusion

This work studies how supervised in-network classification of the rounds of measurements can be adopted to increase the energy efficiency of data collection. In this framework, a sensor network performs classification to transmit to the sink (and then to the base station) only rounds of measurements of interest to the end users. In-network classification requires energy resources, *e.g.* for transmitting features across the network, and can be obviously subject to errors (false positives and false negatives). Thus constraints on the efficiency of data collection imply that the energy available for the distributed feature extraction is bounded by parameters such as the number of hops and the priors of the classes of measurement rounds that are relevant the end users. The approach could be combined with other compression methods for wireless sensor networks (*e.g.* compressed sensing). It could also be useful in scenarios of a very large number of networks (*e.g.* deployed in smart homes) transmitting measurements to the cloud, since it could reduce the amount of data stored and processed at the server and therefore the fees due to the cloud provider.

We study in-network classification in depth via intensity histogram, because the properties of this kind of feature allow

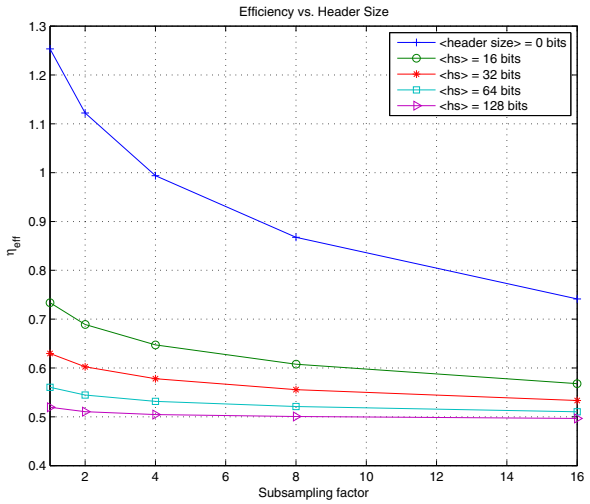


Fig. 11. Effective efficiency for sensing frame classification via $4-$bin histograms.

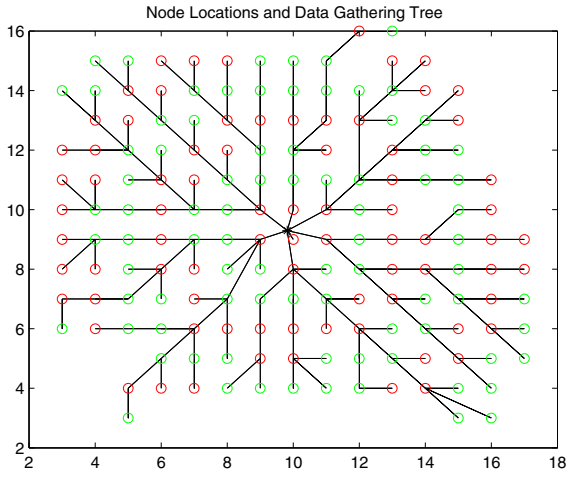


Fig. 12. Sensor network topology for the nodes over the whole Pacific Northwest region: 163 nodes, 6 hops.

to simplify the math of our analysis. The classification cost can be reduced by spatially subsampling the sensor nodes and so by decreasing size of histograms and number of messages necessary to extract them. We present extensive analytical and experimental results for regular, uniform random and Bernoulli subsampling to characterize the classification performance in terms of misclassification rates, mean square error and energy efficiency.

One of our next goals is to find a sensor network dataset provided with independent ground truth and suitable for interesting classification applications, in order to further evaluate this approach. Other possible future directions consist in studying the performance of the framework under a more accurate energy model (*e.g.* including computation costs) and/or for other types of features and classification schemes, integration with compressed sensing and testing over an actual wireless sensor network.

REFERENCES

[1] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. 40, no. 2, April 2008.

[2] J. Predd, S. Kulkarni, and H. Poor, "Distributed learning in wireless sensor networks," *Signal Processing Magazine, IEEE*, vol. 23, no. 4, pp. 56–69, July 2006.

[3] M. Abu Alsheikh, S. Lin, D. Niyato, and H.-P. Tan, "Machine learning in wireless sensor networks: Algorithms, strategies, and applications," *Communications Surveys & Tutorials, IEEE*, vol. 16, no. 4, pp. 1996–2018, 2014.

[4] S. Rajasengar, C. Leckie, and M. Palaniswami, "Anomaly detection in wireless sensor networks," *Topics on Security in Ad Hoc and Sensor Networks*, 8 2008.

[5] J. W. Branch, C. Giannella, B. Szymanski, R. Wolff, and H. Kargupta, "In-network outlier detection in wireless sensor networks," *Knowledge and information systems*, vol. 34, no. 1, pp. 23–54, November 2013.

[6] J. Wang, S. Tang, B. Yin, and X.-Y. Li, "Data gathering in wireless sensor networks through intelligent compressive sensing," in *INFOCOM, 2012 Proceedings IEEE*. IEEE, March 2012.

[7] D. Ebrahimi and C. Assi, "On the benefits of network coding to compressive data gathering in wireless sensor networks," in *Sensing, Communication, and Networking (SECON)*. IEEE, June 2015.

[8] P. Bromiley and N. Thacker, "When less is more: Improvements in medical image segmentation through spatial sub-sampling," in *Proc. MIUA*, 2007.

[9] J.-C. Joo, T.-W. Oh, J.-H. Choi, and H.-K. Lee, "Steganalysis scheme using the difference image of calibrated sub-sampling," in *Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*. IEEE, 2010.

[10] M. Maróti, B. Kusy, G. Simon, and Á. Lédeczi, "The flooding time synchronization protocol," in *Proceedings of the 2nd international conference on Embedded networked sensor systems*. ACM, 2004, pp. 39–49.

[11] L. A. Rossi, B. Krishnamachari, and C.-C. Kuo, "Optimal sink deployment for distributed sensing of spatially nonstationary phenomena," in *Global Telecommunications Conference, 2007. GLOBECOM'07. IEEE*. IEEE, November 2007, pp. 1124–1128.

[12] M. Widmann and C. S. Bretherton, "50 km resolution daily precipitation for the pacific northwest," www.jisao.washington.edu/data_sets/widmann/, University of Washington, Tech. Rep., 1999.

[13] E. J. Candè and M. B. Wakin, "An introduction to compressive sampling," *Signal Processing Magazine, IEEE*, vol. 25, no. 2, pp. 21–30, 2008.