ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

Feature Noise Tuning for Resource Efficient Bayesian Network Classifiers

Laura I. Galindez Olascoaga[1], Jonas Vlasselaer[1], Wannes Meert[2], Marian Verhelst[1]

1- MICAS, Dept. of Electrical Engineering, KU Leuven, Belgium

2- DTAI, Dept. of Computer Science, KU Leuven, Belgium

Abstract. Emerging portable applications require always-on sensing technologies to continuously monitor the environment and their user's needs. Yet, the high power consumption that results from this continuous sensing often hampers these systems' always-on functionality. In this paper we propose a hardware-aware Machine Learning scheme that exploits the devices' ability to trade-off the quality of its sensors versus its power consumption. We introduce a technique that extends Bayesian Network classifiers with hardware description nodes that encode the probabilistic relation between sensory features and their degraded versions. We show how this allows to tune the hardware device's power consumption versus inference accuracy trade-off space with fine granularity, resulting in operating points that achieve significant power savings at almost no accuracy loss. This is empirically shown on various Machine Learning benchmarking datasets.

1 Introduction

Many smart sensory applications (e.g. augmented reality or natural user interfaces in smart watches and phones, domestic gadgets, robots, etc.) encounter a fundamental conflict between the available battery life and the desire to uninterruptedly gather, process and fuse a large amount of high quality sensory information. To overcome limitations on computational bandwidth, state-of-the-art implementations often rely on cloud computing to run the most complex sensor fusion and inference tasks remotely [1]. Yet, this does not address the dominance of the sensor interfaces themselves on the overall power consumption of the device. Moreover, this approach results in a significant overhead from the necessary data transfer on the system's power consumption, as well as increased latency and user privacy concerns [1]. Always-on mobile applications therefore necessitate a new local compute paradigm whereby the algorithmic level of abstraction has knowledge of the hardware's properties and limitations. In this paper we propose to extend Bayesian Network classifiers with nodes that describe the hardware's noise, thus enabling a machine-learning-based resource consumption scalability scheme.

The remainder of this paper is organized as follows. In Section 2, we discuss the state of the art of related research topics. Our noise scalable Bayesian Network classifier and corresponding noise tuning algorithm are presented in Sections 3 and 4, respectively. We experimentally evaluate the proposed methods in Section 5 and conclude in Section 6.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

2 Related Work

Related work on enabling always-on inference in embedded sensing applications is taking place both in the hardware and in the algorithmic research communities. From the hardware design point of view, the development of highly power-efficient processors and sensor front-ends [2, 3], albeit relevant for the realization of the aforementioned paradigm, is often not co-optimized with the algorithms' functionality. The hardware design paradigm addressed in this work, "analog-to-information converters" [4], does open opportunities towards such co-optimization. In this concept, information is already extracted from the incoming sensory signal in the analog domain, which results in a very power-scalable system which can be controlled from the algorithmic level. From the algorithmic point of view, state-of-the-art approaches have focused on 1) sequentially deciding what set of observations provide the most information under scarce resources [5], 2) whether more observations are required to meet the tasks' requirements [6] and 3) cost-aware feature sub-set selection[7]. However, these techniques are often not suitable for multi-sensor time series, where all incoming signals have to be coherently sampled. In addition, they only enable a few coarse operating points in the power versus inference performance trade-off space, as they can only decide to observe a feature or not. In the remainder of this paper, we will propose a methodology that addresses the aforementioned shortcomings to realize an efficient hardware aware embedded sensing paradigm.

3 Noise Scalable Bayesian Network classifier

Sensor front-end hardware allows to scale the quality of the incoming sensory data in exchange for power consumption savings. This can be done by controlling the allowed level of degradation of the sensor front-end — typically the result of the "circuit noise" arising in the sensor itself and the subsequent filters and amplifiers. Under the "analog-to-information converter" hardware design paradigm addressed in this paper, this degradation can be independently controlled for each of the features required by the machine learning algorithm. According to common circuit design practice [8], the power consumption of the hardware components generating each feature (P_i) scales proportionally to the standard deviation s_i they tolerate $P_i = \frac{P_{ref,i}}{2^{2SNR_i}}$, where $P_{ref,i}$ is the power consumption of the feature at the lowest possible noise setting and SNR_i is defined as $\log(1/s_i)$. The total power consumption of the sensor front-end — calculated by adding the power contributions of all features $P_{total} = \sum_i P_i$ — can then be optimally traded-off for a target classification accuracy. To exploit these power saving opportunities and to provide a framework for the aforementioned hardware vs inference performance trade-off, we propose to extend general Bayesian Network classifiers with nodes representing various noisy versions of each feature, such that each of them can be observed at a specific noise tolerance selected from a finite user defined set.

Figure 1(a) shows the proposed model structure of an extended Tree Aug-

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

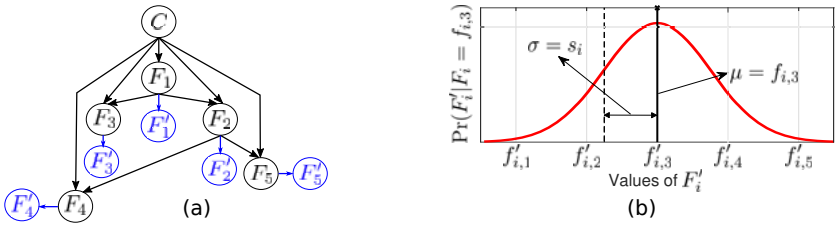

Fig. 1: (a) Example of a noise scalable Bayesian Network classifier — a TAN classifier is extended with noisy nodes (in blue). (b) An example of the probabilistic relation between feature F_i and it's noisy version F'_i.

mented Naive Bayes classifier (TAN) — note that the technique can be implemented with any Bayesian Network classifier. The n-feature classifier has been extended with n additional nodes that encode the probabilistic relations between the features and their noisy versions, denoted by F_i and F'_i, respectively. The noisy version of the feature is described by $F'_i = F_i + Z_i \sim \mathcal{N}(\mu, \sigma)$, where Z_i is the Gaussian circuit noise (conforming to the sensor front-end hardware properties). The mean is the value represented by $\mu = F_i$ and $\sigma = s_i$ is the standard deviation chosen from a user defined set of standard deviations per feature, ordered in increasing size $\mathbf{s_{model}} = \{s_{i,1}, \ldots, s_{i,h}\}$ (see Figure 1(b)). The selection of the possible standard deviations set per feature is made on the basis of the sensor front-end's hardware properties and the tuning capabilities it entails. The proposed structure encodes the following joint probability distribution over the original and noisy feature sets and the class variable C:

$$\Pr(C, F_1, .., F_n, ..., F'_1, .., F'_n) = \prod_{i=1}^{n} \Pr(F'_i|F_i) \cdot \Pr(F_1, \ldots, F_n, C) \ , \qquad (1)$$

where the distribution $\Pr(F_1, \ldots, F_n, C)$ is learned from the training data and the class labels C. The model as such allows to assess the impact that varying the amount of noise s_i on each of its observed features i might have on classification performance. During inference, the nodes corresponding to the noisy feature versions (F'_i in Figure 1) are observed, each with a specific s_i, while the nodes corresponding to the noiseless feature version (F_i) remain always hidden. Features can also be pruned, which is equivalent to observing them with an infinite amount of noise. Given an observation $\{f'_1, f'_2, ..., f'_n\}$, classification is performed by selecting the class that maximizes the posterior probability:

$$\Pr(C|f'_1, f'_2, ..., f'_n) \sim \sum_{F_1} \cdots \sum_{F_n} \prod_{i=1}^{n} \Pr(f'_i|F_i) \cdot Pr(F_1, \ldots, F_n, C) \ . \qquad (2)$$

It is worth mentioning that the last processing block of a real life sensor front-end is an analog-to-digital (ADC) converter. To emulate this behavior, the model described by equation 2 as well as the datasets can be uniformly discretized, as will be described in the experimental Section.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

4 Feature quality scalability

The model described in the previous Section can be used to efficiently choose which of the h noisy versions of each feature i to observe. The objective is to select the set of standard deviations s_i that minimizes the total system power consumption for any given target classification accuracy. This begets a classification performance versus power consumption trade-off space with a variety of feature-wise noise dependent operating points. Due to the computational hardness that the problem of feature subset selection entails [5], we use a greedy search heuristic to determine the Pareto optimal set of such operating points.

We initialize the search to the smallest standard deviation available for all features $\mathbf{s}_{select} = \{s_{1,1}, \ldots, s_{n,1}\}$ and will iteratively perform the following steps until all features are observed with the largest noise setting, i.e. $\mathbf{s}_{select} = \{s_{1,h}, \ldots, s_{n,h}\}$. At each iteration, we target to reduce the quality of each feature individually, thus producing n quality reduction candidates and selecting the best one by means of a greedy neighborhood search: in each candidate j, the noise tolerated by feature F_j is increased one level with respect to the current setting, hence going from s_{j,v_j} to s_{j,v_j+1}, where v_j refers to the current value of s_j. Each of the resulting candidates are described by $\mathbf{s}_{cand,j} = \{s_{1,v_1}, \ldots, s_{j,v_j+1}, \ldots, s_{n,v_n}\}$. To estimate the candidate accuracy (Acc), we inject each feature in the validation set with the corresponding candidate Gaussian noise, and count the number of correctly predicted instances. We then estimate the candidate power consumed across all the features (P_{total}) by summing the individual features' power contributions. We finally select the candidate that minimizes a predefined cost function $CF = \frac{\Delta Acc}{\Delta P_{total}}$, where the term Δ refers to the predicted state difference between time t and time $t+1$.

5 Experimental Evaluation

We evaluated the power-accuracy trade-off enabled by the proposed model for four benchmarking datasets from the UCI Machine Learning repository that correspond to multi-sensor mobile applications and that can benefit from the power consumption scalability aimed at in this paper: 1) mobile robot navigation (Pioneer), 2) human activity recognition from smartphone data (HAR), 3) human activity recognition from body motion and vital signs recordings (MHealth) and 4) physical activity monitoring from an inertial measurement unit and a heart rate monitor (PAMAP). As a pre-processing step, we removed all the nominal features — because we are interested only on sensory signals — and we uniformly discretized the remaining numerical features into 32 bins to emulate the role of a 5 bit ADC in the sensory stream process. We also performed feature selection with Weka's wrapper subset evaluator with a Bayes Net classifier and the default parameters [9] to avoid processing redundant or irrelevant features, a consideration regarding the resource constrains of embedded applications. For all the experiments, we extended Naive Bayes classifiers with seven possible noisy feature settings $\mathbf{s_{model}} = \{0.0005, 0.001, 0.03, 0.05, 0.07, 0.09, \infty\}$ —where

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

Dataset	Instances	Features	Classes	Sel. F.	$P_{max} = 1$	$P_{max}/10$	$P_{max}/100$
Pioneer [1]	6129	27	35	17	95.7 ± 0.2	93.8 ± 2.7	89.5 ± 3.5
HAR	10299	561	6	37	94.8 ± 0.25	94.6 ± 1.1	89 ± 2.5
Mhealth	312475	23	11	21	81.8 ± 0.01	80.4 ± 0.35	75.5 ± 0.4
PAMAP	10000 [2]	39	12	11	88.3 ± 0.001	87.04 ± 0.03	81.9 ± 0.04

Table 1: Experimental data sets' characteristics and the accuracy our methodology achieves for different power scalings.

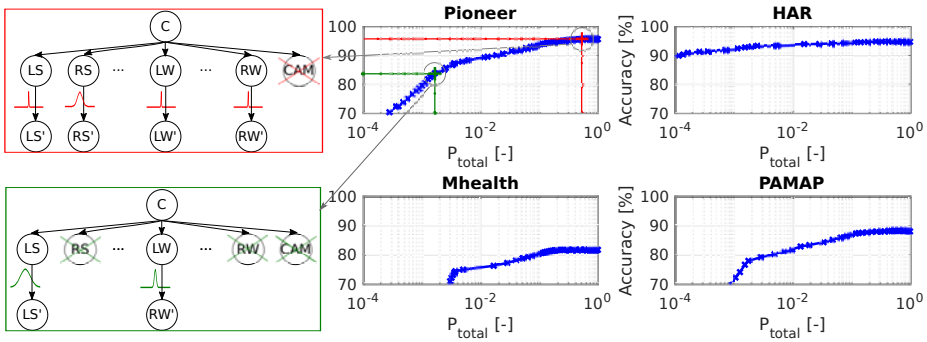

Fig. 2: Right: Power vs accuracy trade-off on all datasets. Left: two operating points from the Pioneer dataset's Pareto optimal front

∞ corresponds to pruning the feature— that were normalized to each feature's dynamic range. We normalized the power consumption estimation from 0 to 1, where $P_{max} = 1$ represents the setting where all features are extracted with their lowest possible noise ($s_i = 0.0005$). This facilitates the performance comparison among benchmarks, given that the dataset generating hardware is unknown. We performed a 10-trial, 5-fold cross validation of the greedy search detailed in Section 3 to asses the applications' power consumption vs accuracy trade-off, as reflected by the Pareto curves of Figure 2. All datasets demonstrate power saving opportunities of at least an order of magnitude whilst preventing accuracy from degrading more than 2%. Further power savings can be traded off when accuracy requirements are relaxed. For example, consider the Pioneer data set (Figure 2) whose features were extracted from a variety of sensors among which are sonars (LS,RS), wheel odometers (LW,RW) and a multi-channel camera (CAM). Accuracy degradation can be avoided for power consumption savings of more than 3× (top-left) but savings of up to 3 orders of magnitude (bottom-left) can be achieved if the accuracy requirements decrease from 95% to 83%.

The last three columns of Table 1 show the accuracy attained by the selection strategy at different power scaling levels of the sensor system for all the datasets. Accuracy degradation rates depend on the data sets' properties: number of in-

[1] From movement experiences only
[2] Reduced data set size used for experiments

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

stances, classes and features; number and position of the discretization intervals; and the class conditional probability distribution of each feature. Overall, all the data sets benefit from the noise tuning algorithm: in all our experiments, they lose less than 10% accuracy while achieving sensor power savings of at least two orders of magnitude. This also proves that the methodology can be effectively implemented with a wide variety of applications as it allows to select the optimal settings according to the dataset and the classification task.

6 Conclusion and future work

In this paper we proposed to extend Bayesian Network classifiers with parametrizable hardware description nodes that encode the probabilistic relation between sensory features and their degraded versions. We demonstrated that this model allows to exploit the power saving opportunities of feature tunable sensing systems by allowing to optimally tune the noise across sensory features. We discussed the performance of the proposed methodology through the analysis of the achievable trade-off between power consumption and inference accuracy, and we demonstrated the general applicability on four standard Machine Learning datasets relevant to embedded sensing applications. Since this scheme allows to scale the hardware power consumption from the algorithmic level of abstraction, it can form the basis for a number of run-time strategies that can be implemented in embedded devices to ensure their always-on functionality.

References

[1] F. H. Bijarbooneh, W. Du, E. C. H. Ngai, X. Fu, and J. Liu. Cloud-assisted data fusion and sensor selection for internet of things. *IEEE Internet of Things Journal*, 3(3):257–268, 2016.

[2] B. Moons and M. Verhelst. Energy-efficiency and accuracy of stochastic computing circuits in emerging technologies. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 4(4):475–486, 2014.

[3] F. Chen, A. P. Chandrakasan, and V. M. Stojanovic. Design and analysis of a hardware-efficient compressed sensing architecture for data compression in wireless sensors. *IEEE Journal of Solid-State Circuits*, 47(3):744–756, 2012.

[4] M. Verhelst and A. Bahai. Where analog meets digital: Analog to information conversion and beyond. *IEEE Solid-State Circuits Magazine*, 7(3):67–80, 2015.

[5] Andreas Krause and Carlos Guestrin. Optimal nonmyopic value of information in graphical models: efficient algorithms and theoretical limits. 2005.

[6] Arthur Choi, Yexiang Xue, and Adnan Darwiche. Same-decision probability: A confidence measure for threshold-based decisions. *International Journal of Approximate Reasoning*, 53(9):1415–1428, 2012.

[7] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge L Reyes-Ortiz. Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine. In *International workshop on ambient assisted living*, pages 216–223. Springer, 2012.

[8] Rahul Sarpeshkar. Analog versus digital: extrapolating from electronics to neurobiology. *Neural computation*, 10(7):1601–1638, 1998.

[9] Ron Kohavi and George H John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2):273–324, 1997.