POWER DISSIPATION LIMITS AND LARGE MARGIN IN WIRELESS SENSORS

Shantanu Chakrabartty and Gert Cauwenberghs

ECE Department, Johns Hopkins University, Baltimore, MD 21218, USA
E-mail: {*shantanu,gert*}@*jhu.edu*

ABSTRACT

Wireless smart sensors impose severe power constraints that call for power budget optimization at all levels in the design hierarchy. We elucidate a connection between statistical learning theory and rate distortion theory that allows to operate a wireless sensor array at fundamental limits of power dissipation. $Gini$SVM, a support vector machine kernel-based classifier based on quadratic entropy, is shown to encode the sensor data with maximum fidelity for a given constraint on transmission budget. The transmission power is minimized by $Gini$SVM in the form of a quadratic cost function under linear constraints. A classifier architecture that implements these principles is presented.

1. INTRODUCTION

Emerging wireless embedded sensors impose serious restrictions on power consumption [1]. These constraints derive from the vision of operating the sensors off ambient (*e.g.*, solar or thermal) power, and make it necessary to allocate power resources efficiently to the task of sensing, computation and communication.

As physical size decreases so does energy capacity. Because communication is often the single largest energy consumer the optimization of wireless communication protocols is key in meeting energy constraints. For a typical bluetooth application [2] operating at $700kbps$ and dispensing $115nJ/bit$, the total power consumption of a stand-alone communication module is approximately $800mW$ which exceeds the power budget for most miniature stand-alone sensors. To achieve sub-microwatt operation, power budget optimization has to be performed rigorously at several levels [3] as shown in increasing level in the design hierarchy:

1. **Technology Level**— Supply and threshold voltage reduction;
2. **Circuit and Logic Level**— Logic style, current starvation, switching behavior, supply switching and subthreshold design;
3. **Architectural Level**— Parallelism and pipelining;
4. **Algorithmic and System Level**— Utilization of signal statistics, floor planning, data encoding, sleep modes and reference localization.

This research is supported by a grant from The Catalyst Foundation, New York.

It has been argued [3] that higher reduction in power consumption can be obtained by optimizing at the highest level. Such a hierarchical optimization methodology is a two-way procedure where each level imposes design constraints on the subsequent higher level. For wireless smart sensor arrays, we approach the problem by addressing power-efficient classification and encoding of sensor data transmitted over a communication channel. Considerations from the perspective of low-power sensor design are:

1. **Reduction of Supply Voltages** leading to reduction in noise margins. Noise in the circuit plays a dominant role in the classifier performance, and large margin machine learning techniques may be employed during training to aid in discriminating between classes in run-time.
2. **Reduction of Transmission Rate**, an effective way to reduce power dissipation in the transmission module. Effective reduction of transmission rate amounts to efficient data encoding such that only relevant events are detected and transmitted. For a sensor the primary aim is detection, therefore a classifier is required to register detected events as illustrated in Figure 1. For a multi-event (*e.g.*, olfactory) sensor array architecture it is often necessary to transmit confidence values of events rather than indicator flags to enhance resolution at the receiver, using a more a complex decoding algorithm. With a reduced supply voltage a multi-event/class classifier has to deal with reduced noise margins and increased distortion. In designing a classifier for a low-power wireless sensor array, it is therefore crucial to encode events efficiently according to confidence values to achieve maximum fidelity given the constraint on the number of bits for encoding.

Rate Distortion Theory provides lower bounds on distortion for a fixed transmission rate, determined by the

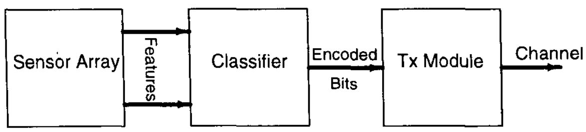

Fig. 1. *Wireless smart sensor architecture*

transmitter power budget. *Statistical Learning Theory* on the other hand provides tools to design classifiers that model signal statistics for efficient discrimination between classes. In this work we elucidate a connection between the two that provides a unified framework to a sensor and classification architecture that operates at fundamental limits of power dissipation.

The paper is organized as follows. Section II briefly reviews principles of rate distortion functions and fundamental limits of distortion given a fixed transmission rate. Section III reviews principles of large margin classifiers and statistical learning theory and discusses how the large margin concept is directly related to designing classifier with larger noise margins for low voltage operation. Section IV presents $Gini$SVM, a large margin classifier that conforms to principles of rate distortion theory to operate at fundamental limits of power dissipation. Section V discusses the practical implications and architectural implementation of $Gini$SVM. Section VI provides comments and conclusions.

2. RATE DISTORTION THEORY

Classical Rate Distortion Theory [4] specifies a lower bound on the number of bits required to encode a signal within a specified amount of distortion. Given a set of independent signals $X_1, X_2, .., X_N$, with power $S_1, S_2, .., S_N$, the optimal allocation of number of bits $R_1, R_2, .., R_N$ to represent $X_1, X_2, .., X_N$ such as to keep the total distortion $D_1 + D_2 + .. + D_N$ below an upper-bound D, is given by

$$\begin{aligned} R &= \sum_i R_i \\ &= \sum_i \frac{1}{2}(\log S_i - \log D_i) \qquad (1) \\ &= \sum_i \frac{1}{2}[\log S_i - \log D_{\min}]_+ \end{aligned}$$

where $[x]_+$ represents positive part of x. Distortion for each of the signals corresponding to this optimum bit allocation has the form

$$D_i = \begin{cases} D_{\min} & ; \quad S_i \geq D_{\min} \\ S_i & ; \quad S_i < D_{\min} \end{cases} \qquad (2)$$

where $D_{\min}$ is computed by solving $\sum_i D_i = D$ through the reverse water-filling procedure illustrated in Figure 2 [9].

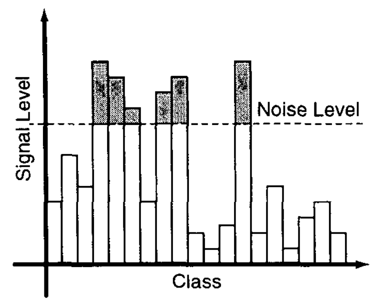

Fig. 2. *Reverse water-filling procedure: signal classes above the noise floor are encoded for transmission.*

For designing integrated sensors an alternative equivalent problem formulation is of more importance. Given the total number of bits R, specified by the transmitted power budget, the alternative seeks to optimize bit allocation amongst various signal to achieve minimum distortion in a mean square sense. The sensor utilizes the optimal allocation obtained to encode the confidence/probability measure for each of the events/classes. Current techniques use dynamic programming to solve the problem of resource allocation. We will show in the next section that by using a large margin classification architecture based on quadratic entropy, $Gini$SVM, an optimal encoding is obtained under the constraint of utilizing a fixed number of bits.

3. LARGE MARGIN CLASSIFICATION

Large margin (LM) classifiers like *Support Vector Machines* (SVM) [5] have several attractive properties from a practical implementation perspective:

1. They generalize well even with relatively few data points in the training set, and bounds on the generalization error can be directly estimated from the training data. This ensures shorter time for real-time system training.

2. The only parameter that needs to be chosen is a penalty term for mis-classification which acts as a regularizer [6] and determines a trade-off between resolution and generalization performance, to control learning ability.

3. The algorithm finds, under general conditions, a unique separating decision surface that provides best out-of-sample performance. This property is unlike neural network classifier implementation where the solution obtained is not unique and hence cannot be quantified.

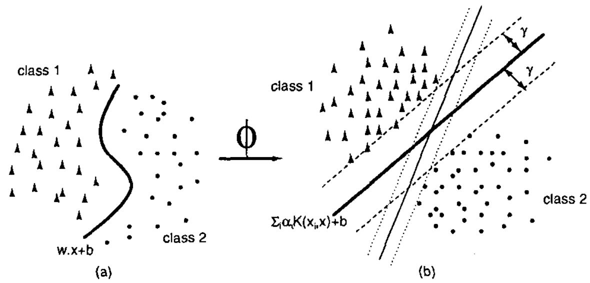

Fig. 3. *Large margin kernel machines.* (a) *Training data in data space.* (b) *Nonlinear map* $\Phi()$ *projects data space onto feature space where they are linearly separated with maximum margin.*

4. They provide a framework to model non-linear classification boundaries by projecting the input data point into higher dimensional space and then computing the distances with the aid of a kernel. This is illustrated in figure 3 where the training samples that are not linearly separable in *(a)* are projected through a non-linear mapping $\Phi()$ onto higher dimensional feature space *(b)* where they become linearly separable.

5. The learning algorithm performs model selection based on some optimization criterion, by which only the data points (support vectors) which are relevant to the classification problem are used for computation. This feature leads to optimal utilization of hardware resources (memory to store support vectors) which is critical for a power conscious design.

One attractive property of large margin classifiers from an implementation perspective, is the direct correspondence of classifier margin maximization during training to optimizing circuit noise margin. This is illustrated in figure 3(b) which shows a maximal margin hyper-plane obtained after training. The maximum margin hyper-plane is more robust to small perturbations in training samples than any other separating hyperplane, and hence is more immune to sensor noise inherent in circuit design. Traditional implementation of non-linear classifiers using (unregularized) neural networks primarily aims at separating the data without direct consideration of margin leading to less optimal noise immunity from a circuit designer perspective. Another attractive property of a large margin classifier is the relative insensitivity of classification performance to small perturbations in the non-linear mapping Φ. This enables use of simple computational elements in an array configured so it can be characterized via a kernel and optimally trained to maximize its classification performance and power consumption.

*Gini*SVM [7] is a sparse multi-class probability regression technique based on large margin principles that generates conditional probability estimates for class k based on input feature vector $\mathbf{x}$, $P_k(\mathbf{x}) \propto exp(f_i(\mathbf{x}))$. The functions $f_i(\mathbf{x})$ are estimated using empirical data consisting of training examples $\mathbf{x}[m], m = 1,..,M$ and their corresponding labels indicating prior probabilities values $y_i[m], i = 1,..,K$. As with SVMs, dot products in the expression for $f_i(\mathbf{x})$ convert into kernel expansions over the training data $\mathbf{x}[m]$ by transforming the data to feature space [8]

$$\begin{aligned} f_i(\mathbf{x}) &= \mathbf{w}_i.\mathbf{x} + b_i \\ &= \sum_m \lambda_i^m \, \mathbf{x}[m].\mathbf{x} + b_i \\ &\xrightarrow{\Phi(\cdot)} \sum_m \lambda_i^m \, K(\mathbf{x}[m], \mathbf{x}) + b_i \end{aligned} \tag{3}$$

where $K(\cdot,\cdot)$ denotes any symmetric positive-definite kernel[1] that satisfies the Mercer condition, such as a Gaussian radial basis function or a polynomial spline [6].

The parameters λ_i^m in (3) are determined by minimizing a dual formulation which for *Gini*SVM takes the form of a quadratic-entropy based potential function in the parameters [7]

$$H_g = \sum_i^M [\frac{1}{2}\sum_l^N \sum_m^N \lambda_i^l Q_{lm} \lambda_i^m + \gamma C \sum_m^N (y_i[m] - \lambda_i^m/C)^2] \tag{4}$$

subject to constraints

$$\sum_m \lambda_i^m = 0 \tag{5}$$

$$\sum_i \lambda_i^m = 0 \tag{6}$$

$$\lambda_i^m \leq C y_i[m] \tag{7}$$

where $Q_{lm} = K(\mathbf{x}[l], \mathbf{x}[m])$.

Using the first order conditions for optimizing (4) the following necessary and sufficient condition is obtained

$$\sum_i^M [f_i(x) - z]_+ = 2\gamma \tag{8}$$

which conforms to the rate distortion criterion (2). The procedure finds the optimal number of bits to encode the signal $exp(f_i(x))$ under the assumption that the log-signal power $\log S_i$ is linearly proportional to the signal strength $f_i(x)$. The parameter 2γ determines the total bit budget obtained from power consumption requirement of the transmitter module. The parameter z is adaptively computed for each classification based on the *reverse water filling*

[1] $K(\mathbf{x},\mathbf{y}) = \Phi(\mathbf{x}).\Phi(\mathbf{y})$. The map $\Phi(\cdot)$ need not be computed explicitly, as it only appears in inner-product form.

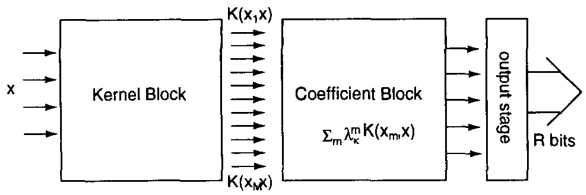

Fig. 4. *GiniSVM Architecture*

principle illustrated in figure 2. Only signals with power higher than z or the noise level are encoded for transmission and the rest are discarded. The total bit budget 2γ is then proportionately distributed according the class bit budget $R_i = [f_i(x) - z]_+$. The validity of signal independence $f_i(x)$ between K classes is obtained through training of $Gini$SVM which makes $f_i(x)$ depend only on class discriminatory features.

4. SYSTEM ARCHITECTURE

The architecture of $Gini$SVM is shown in figure 4. It consists of the following blocks:

- **Kernel Block**— stores support vectors $\mathbf{x}[m]$ and computes the kernel $K(\mathbf{x}[m], \mathbf{x})$ between support vectors and the input feature vector $\mathbf{x}$. Floating gate transistor array [10] could be a possible implementation ensuring a minimal standby power consumption.

- **Coefficient Block**— computes the inner-product between the kernel $K(\mathbf{x}[m], \mathbf{x})$ and the coefficients λ_i^m to obtain values of $f_k(\mathbf{x}) = \sum_m \lambda_i^m K(\mathbf{x}[m], \mathbf{x})$.

- **Encoder Output Block**— encodes confidence values of $P_i(\mathbf{x}) = exp(f_i(\mathbf{x})), i = 1, .., K$ using a total fixed number of bits R.

An efficient analog implementation of the kernel and coefficient block can be obtained using subthreshold CMOS circuits as described in [10].

5. CONCLUSIONS

In this paper we proposed a classifier architecture for a wireless sensor operating at fundamental limits of power dissipation by utilizing the $Gini$SVM large-margin kernel machine. The design methodology is flexible enough to incorporate user specified bit budget constraints into the classifier architecture. Circuit design parameters are then directly obtained by optimizing a classifier cost function which also ensures that the system operates with a near optimal fidelity.

6. REFERENCES

[1] Chandrakasan, A. et. al, "Power Aware Wireless Microsensor Systems," *Keynote Paper ESSCIRC*, Florence, Italy, 2002.

[2] The Official Bluetooth Wireless Info Site, *http://www.bluetooth.com*

[3] Shanbhag, N.R, "A Mathematical Basis for Power-Reduction in Digital VLSI Systems," *IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing,* vol. **44**, pp. 935-951, 1997.

[4] Shannon C.E, "A Mathematical Theory of Communication," *The Bell System Technical Journal,* vol. **27**, pp. 379-423,623-656, 1948.

[5] Vapnik, V. *The Nature of Statistical Learning Theory,* New York: Springer-Verlag, 1995.

[6] Girosi, F., Jones, M. and Poggio, T. "Regularization Theory and Neural Networks Architectures," *Neural Computation,* vol. **7**, pp 219-269, 1995.

[7] Chakrabartty, S. and Cauwenberghs, G. "Forward Decoding Kernel Machines: A hybrid HMM/SVM Approach to Sequence Recognition," *IEEE Int. Conf. of Pattern Recognition: SVM workshop. (ICPR'2002),* Niagara Falls, 2002.

[8] Schölkopf, B., Burges, C. and Smola, A., Eds., *Advances in Kernel Methods-Support Vector Learning,* MIT Press, Cambridge, 1998.

[9] Cover T.A, Thomas J.A, *Elements of Information Theory,* John Wiley and Sons, 1991.

[10] Chakrabartty, S., Singh, G. and Cauwenberghs, G. "Hybrid Support vector Machine/Hidden Markov Model Approach for Continuous Speech recognition," *Proc. IEEE Midwest Symp. Circuits and Systems (MWSCAS'2000),* Lansing, MI, Aug. 2000.