A NEW APPROACH TO ENERGY EFFICIENT CLASSIFICATION WITH MULTIPLE SENSORS BASED ON ORDERED TRANSMISSIONS

Rick S. Blum, Yusuf Artan

Lehigh University
Bethlehem, PA 18015
rblum@eecs.lehigh.edu, yua206@lehigh.edu

Brian M. Sadler

Army Research Laboratory
Adelphi, MD 20783
bsadler@arl.army.mil

ABSTRACT

Classification employing sensors connected by wireless networks is of great interest. As the sensor nodes are usually powered by batteries, saving transmissions is important. We demonstrate transmissions can be saved, without degradation in error probability, using an ordering approach. The average number of transmissions saved (ANTS) is lower bounded by a quantity proportional to the number of sensors employed provided a well-behaved distance measure between the sensor distributions is sufficiently large. For such cases, the ANTS over the optimum unconstrained energy approach is shown to be larger than half the number of sensors employed.

Index Terms— energy efficient classification, sensor networking, cross-layer design, joint signal processing and communications, pattern recognition

1. INTRODUCTION

Classification has been extensively studied from classical perspectives [1],[2] which ignore energy consumption. More recently, small sensor equipped nodes, called sensors here, carrying their own energy sources and using wireless communication have been of great interest for use in classification under the topic of sensor networking. As these nodes carry their own batteries, efficient use of their battery power implies longevity of the sensor networks. Therefore, there is a need to develop an energy efficient classification system that has a longer lifetime while preserving the error probability of the classifier.

Consider a classification approach which collects scalar data samples from N sensors which are transmitted by wireless links to a fusion center and formed into an N-dimensional vector which is used to make a binary decision at the fusion center. Such a classification approach uses a training procedure to separate the N-dimensional space, formed from the N-dimensional vectors of possible sensor observations, into two distinct regions. Then, after training, depending of which of the two distinct regions the current N-dimensional vector under test falls into, this classification approach will decide for one of the two hypothesis. We call this approach, where all sensors transmit their observations to the fusion center, the energy unconstrained approach.

Let the data sample taken from the j^{th} sensor be denoted by x_j. Further, let us focus on classification approaches which are linear in the features, where the the features are some nonlinear processing $g_j(x_j)$ of the j^{th} sensor's observations. Thus, the unconstrained energy approach will decide for H_1 when

$$\sum_{j=1}^{N} L_j = \sum_{j=1}^{N} w_j g_j(x_j) > \beta \tag{1}$$

and it will decide for H_0 otherwise, where the weight vector $(w_1, \ldots, w_N)^T$ is found during training along with the threshold β which allows a bias towards one hypothesis or the other ($p = Prob(H_1) \neq 0.5$). This formulation applies to all classification approaches which are linear in the feature space including support vector machine (SVM) networks [3],[4],[5], perceptron networks [6],[7], potential function networks [8], and radial basis function networks [9], just to name a few. It is worth noting that networks that make decisions using (1) have been shown to have universal approximation capabilities which makes these networks of considerable practical interest. In the sequel, we denote $y_j = g_j(x_j)$ for brevity.

We assume the wireless links are reliable. Since the sensors are assumed to be close to one another, we assume they all use the same energy to send data to the fusion center. We can extend our results to other cases in a straightforward manner. Now we discuss a method for saving energy based on using ordered transmissions. Assume the same training procedure as for the energy unconstrained approach is undertaken. This establishes the hyperplane in terms of $y_1, \ldots, y_N$ described in (1). Recall, any hyperplane is defined by a vector perpendicular to the hyperplane, assuming the vector's originating endpoint falls within the hyperplane. In (1), $(w_1, \ldots, w_N)$ is such a vector. We take the vector's originating endpoint as the origin for simplicity. Now we can

THIS MATERIAL IS BASED ON RESEARCH PARTIALLY SUPPORTED BY THE AIR FORCE RESEARCH LABORATORY UNDER AGREEMENT NO. NO. FA9550-06-1-0041, AND BY THE ARMY RESEARCH LABORATORY.

define the informativeness value of any observation vector by the length of it's projection onto the vector perpendicular to the hyperplane. Thus the magnitude of the informativeness is roughly speaking the shortest distance from the point $(y_1, \ldots, y_N)$ to the hyperplane. A positive informativeness value indicates a preference for one hypothesis (H_1 in this paper without loss of generality) and a negative value indicates a preference for the other (H_0 in this paper). The magnitude indicates how much evidence for the corresponding hypothesis the given data value indicates.

In order to define the informativeness of a sensor observation, each of the sensor observations can be viewed as a vector if we set all other components, but the one corresponding to the observation in question, to zero. Thus the vector corresponding to second observation, assumed to be y_2, is $(0, y_2, 0, \cdots, 0)^T$. Then we can order the transmissions so the more informative (larger magnitude projections) sensors transmit first and the transmissions are stopped after overwhelming evidence is accumulated for one hypothesis or the other. Note that the evidence is accumulated in the form of the sum of the informativeness values for sensors that transmitted. Using the method just described, we calculate the informativeness value of the observation corresponding to y_j as $L_j = w_j y_j$ and we order these. Recall the sensor transmissions are wireless so sufficiently close nodes, as we assume, can hear each others' transmissions. Thus, for sufficiently close nodes, all nodes can listen and compute the sum of the informativeness values that have been transmitted[1]. When they judge this sum to be large or small enough, they stop transmitting, which will save energy. We can show that a proper approach of the type we suggest will outperform, employ less sensor data transmissions for the same error probability, the optimum unconstrained energy approach.

To order transmissions in a distributed manner, not requiring any coordination transmissions between the sensors, the j^{th} sensor can transmit at a time equal to $\frac{K}{|L_j|}$ for some real scalar K that can be chosen as small as desired, within system limits, to minimize any delays. All sensors (or fusion center) listen and compute the sum of all the L_js transmitted. Then the transmissions stop when this sum is larger than a threshold t_U or smaller than a threshold t_L which we now define. Let n_{UT} be the number of sensors who have not yet transmitted at a given time and let $L_{N-n_{UT}}$ denote the informativeness value for the last sensor transmission prior to that same time. Then

$$t_U = \beta + n_{UT}|(L_{N-n_{UT}})| \tag{2}$$

and

$$t_L = \beta - n_{UT}|L_{N-n_{UT}})|. \tag{3}$$

Since we need to order the informativeness values to order transmissions, we employ the notation $|L_{[1]}| > |L_{[2]}| > \cdots > |L_{[N]}|$ for the ordered values. Thus $L_{[1]}$ denotes the informativeness value with largest magnitude and $L_{[2]}$ has the next largest magnitude. We will show the energy saving that can be achieved with this approach by counting the transmissions saved.

[1]Alternatively the fusion center can send a control message for sensors to stop transmissions so sensor listening is avoided.

2. THEORETICAL ANALYSIS OF PERFORMANCE ADVANTAGE OF ORDERING TRANSMISSIONS

The following theorem demonstrates that the general approach of having more informative sensors transmit first will save transmissions.

Theorem 1 *Assume sensor transmissions are ordered so sensors with larger $|L_i|$ transmit first and that transmissions stop when the left-hand side of (1) is larger than t_U from (2) or smaller than t_L from (3) for the given sensor transmissions. If t_U is exceeded we decide for H_1. If t_L is exceeded we decide for H_0. If all transmissions proceed we use (1). In (1), $(w_1, \ldots, w_N)^T$ (used in $L_i = w_i y_i$) and β are obtained from training for an energy unconstrained approach (no restrictions on energy). The approach described in this theorem is always better than the energy unconstrained approach. It gives the same probability of error p_e while using a smaller average number of sensor data transmissions.*

Outline of the Proof for Theorem 1 *Note that we have ordered the transmissions. Thus if the last transmission sent the sensor informativeness value $L_{N-n_{UT}}$, then the largest possible magnitude contribution from the sum of the sensor informativeness values not yet transmitted is $n_{UT}|L_{N-n_{UT}}|$. Note that t_U is chosen to be the unconstrained energy threshold β plus this extra safety margin. Once the sum of the informativeness values from the transmitting sensors is larger than t_U, this implies the fusion center's sum, from (1), will have to be larger than β, regardless of the data observed at the sensors that did not transmit. Thus, even without transmitting further, for this set of observations we are able to implement the energy unconstrained approach with fewer transmissions. A similar savings of transmissions can be made when the sum of the informativeness values from the transmitting sensors is smaller than t_L, which implies the fusion center's informativeness sum will have to be smaller than β, regardless of the data observed at the sensors that did not transmit.*

So far, we have shown that an approach based on sending signals from the different sensors at different times can be more efficient. Next, we show large gains under fairly mild conditions.

Consider a hypothesis testing problem with a corresponding distance measure [10] whose value s measures the distance between the distributions of the sensor observations occurring under the two hypotheses. Thus if s becomes large it should be easy to decide the true hypothesis based on L_i for any i. The sign of L_i should be consistent with the true hypothesis as described previously. Of course H_0 should lead to

$L_i < 0$ in a similar way. Let us quantify this in the following, very reasonable, assumption.

Assumption 1 *For the binary hypothesis testing problems considered, we assume the existence of a distance measure with value s such that $Pr(L_j > 0|H_1) \to 1$ as $s \to \infty$ and $Pr(L_j < 0|H_0) \to 1$ as $s \to \infty$.*

In the numerical results, we give an example of such a binary hypothesis testing problem which we call the shift-in-mean problem.

Let us consider the number of transmissions saved by the approach considered in Theorem 1. One question of particular interest is how the savings scale with the number of sensors employed N. The following theorem compares to the optimum unconstrained energy approach.

Theorem 2 *When using the approach in Theorem 1 for a binary hypothesis testing problem satisfying Assumption 1 with sufficiently large s, the average number of transmissions saved over the optimum unconstrained energy approach increases proportional to N (for even N) while error probability is not effected. In particular, the average number of transmissions saved N_s is lower bounded by $\lfloor \frac{N}{2} \rfloor$ for large N.*

Outline of the Proof for Theorem 2 *First assume $\beta = 0$. We return to $\beta \neq 0$ later. Let us focus on the case where the upper threshold is exceeded and N is even. Define*

$$k^* = min_{1 \leq k \leq N} \left\{ \sum_{i=1}^{k} L_{[i]} > \beta + (N-k) \left| L_{[k]} \right| \right\} \quad (4)$$

as the number of necessary transmissions, so that the average number of transmissions saved N_s is $E\{N - k^*\}$. It is easy to show that (more saved if we consider lower threshold)

$$\begin{aligned} N_s &\geq E\{N-k^*\} = \sum_{k=1}^{N} (N-k) Pr(k^* = k) \\ &\geq \frac{N}{2} Pr(k^* \leq \frac{N}{2}) \end{aligned} \quad (5)$$

since $\sum_{k=1}^{N} (N-k) Pr(k^* = k) \geq \sum_{k=1}^{N/2-1} \frac{N}{2} Pr(k^* = k) + \frac{N}{2} Pr(k^* = \frac{N}{2})$ by dropping positive terms and using $N - k > N/2$ if $k < N/2$.

Now consider $Pr(k^* \leq \frac{N}{2})$ from (5) with k^* from (4). Intuitively, when $k = N/2$ on the right hand side of (4) for even N, each term in the sum on the left hand side of the inequality is individually larger in magnitude than the multiplier of the term $(N-k) = \frac{N}{2}$ on the right hand side. Thus, if the terms in the sum are all positive, the value of the sum is larger than $\frac{N}{2} \left| L_{[\frac{N}{2}]} \right|$. Under H_1 and for sufficiently large s, the terms in the sum all have a positive sign from Assumptions 1.

Thus we have shown that $Pr(k^* \leq \frac{N}{2}) \to 1$ as $s \to \infty$ under H_1 from Assumptions 1. Thus if $p = 0.5$, H_1 is true half the time and so $N_s \geq N/4$ from (5). We can bound the gains from exceeding the lower threshold under H_0 with an identical factor of $N/4$ and these gains add to give $N_s \geq N/2$.

Now consider $\beta \neq 0$. If we consider (4), we see that the first and last term inside the $\{\}$ grow with N, while the bias term β does not. Thus as N increases the bias term must become less and less important. Thus, at some large N the importance of this term becomes negligible and the previous conclusions still apply.

3. NUMERICAL RESULTS

Here, we focus on a shift-in-mean problem where at the j^{th} sensor we observe

$$x_j = \theta - s/2 + n_j \quad (6)$$

where $\theta = 0$ under H_0 and $\theta = s > 0$ under[2] H_1. In (6), n_j represents the noise which has probability density function (pdf) f with zero mean and unit variance. To promote simplicity we focus on linear classification approaches so that $y_i = x_i$. The difficulty of the hypothesis testing problem described in (6) is clearly characterized by the signal strength s. Clearly as s becomes very large, y_j yields large positive values under H_1 and large negative values under H_0 with very high probability. Without loss of generality, we assume that the training will preserve this polarity so that in cases of large s, $\sum_{j=1}^{N} L_j = \sum_{j=1}^{N} w_j y_j$ yields large positive values under H_1 and large negative values under H_0 with very high probability. This implies positive w_j for this formulation and this would come out of sufficiently accurate training. Choosing this formulation, where positive values correspond to H_1 and negative to H_0 is not limiting and other choices would yield similar results. The main point is that the two hypothesis are easier to separate as s becomes large as we show in the following Lemma.

Lemma 1 *For the shift-in-mean problem, as described in (6), $Pr(L_j > 0|H_1) \to 1$ as $s \to \infty$ assuming $w_j > 0$.*

Outline of the Proof for Theorem 3 *Given $w_j > 0$ we have*

$$\begin{aligned} Pr(L_j > 0|H_1) &= Pr(w_j(\frac{s}{2} + N_j) > 0|H_1) \\ &= Pr(\frac{s}{2} > -N_j|H_1) = Pr(-\frac{s}{2} < N_j|H_1) \\ &= \int_{y_j=-s/2}^{\infty} f(y_j) dy_j = 1 - F(-\frac{s}{2}) \end{aligned} \quad (7)$$

where F is the cumulative distribution function (cdf) corresponding to f. Now since $F(-\infty) = 0$ for any cdf, then from (7) we obtain

$$\lim_{s \to \infty} Pr(L_j > 0|H_1) = \lim_{s \to \infty} 1 - F(-\frac{s}{2}) = 1 \quad (8)$$

[2]We assume s takes on a specific value.

Similar demonstration for $Pr(L_j < 0|H_0)$ follows easily.

First consider the model in (6) with unit variance additive Gaussian noise and assume that either hypothesis is equally likely ($p = 0.5$) and that a SVM is used for the classification (500 training samples, and 2500 testing samples) . Figure 1 shows the percentage of transmissions saved by our approach for cases with signal strengths of $s = 0.5$, $s = 1$, $s = 5$ plotted as a function of the total number of sensors in the network. From the figure, we can observe the benefits of having larger signal strength on the performance. Monte Carlo simulation results which use 1000 runs were employed to obtain these results. As the signal strength is increased from $s = 1$ to $s = 5$, the number of saved transmissions increases. Our results indicate that savings converge to an upper bound, which is close to $N_s = 0.56N$ for large N and s. For smaller signal strengths (for example: $s = 0.5$), Figure 1 indicates that our savings are under $0.5N$,which is consistent with Theorem 2. The results in Figure 2, for $s = 5$, show that the savings are

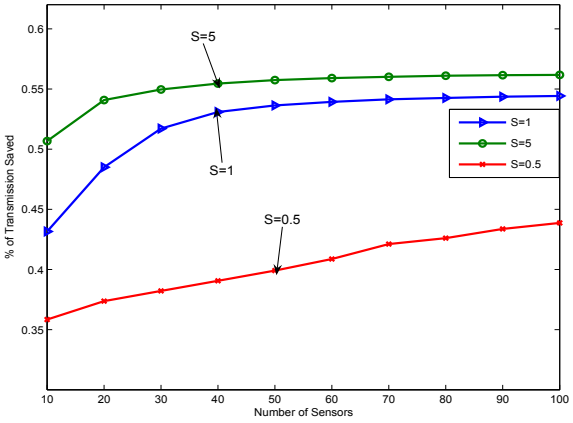

Fig. 1. Percentage of transmissions saved over unconstrained support vector machine classifier for $p = 0.5$ and sensor observations from (6) with unit variance additive Gaussian noise.

larger for $p \neq 0.5$ than for $p = 0.5$. Further, the savings for $p < 0.5$ appear to be same as those for $p > 0.5$ provided that $|p - 0.5|$ is the same in both cases. We have many more numerical results which we omit due to space. The results show that p_e for our ordering approach is the same as p_e for the energy unconstrained approach, as expected. We also find results similar to those in Figure 1 and Figure 2 for noise with a beta, gamma or uniform distribution.

4. CONCLUSION

In this work, we describe a new approach for saving transmissions using ordering for a classification problem with multiple sensors.

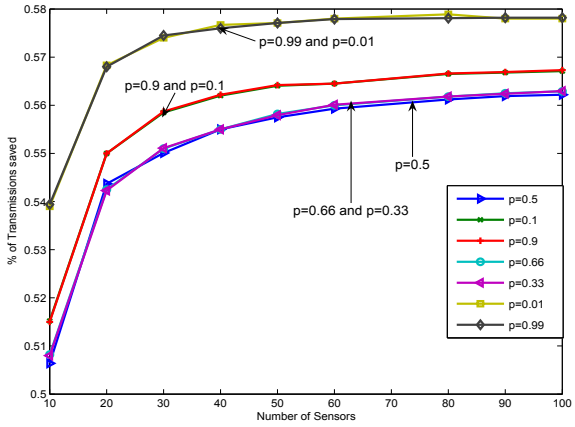

Fig. 2. Percentage of transmissions saved over unconstrained support vector machine classifier for various $p \neq 0.5$, $s = 5$, and sensor observations from (6) with unit variance additive Gaussian noise.

5. REFERENCES

[1] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd Edition, Wiley, 2001.

[2] S. Theodoridis, K. Koutrambas, *Pattern Recognition*, 2nd Edition, AP, 1998.

[3] C. Cortes and V. Vapnik, Support-Vector Networks, Machine Learning, 20, 1995

[4] S. Abe, *Support Vector Machines for Pattern Recognition*, Springer, New York, 1st Edition, 2005.

[5] V. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1st Edition, 1998.

[6] F. Rosenblatt "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain" , Psychological Review, v65, No. 6, pp. 386-408, 1958

[7] S. I. Gallant, "Perceptron-based learning algorithms" IEEE Transactions on Neural Networks, vol. 1, no. 2, pp. 179-191, 1990

[8] S. Lee and R. M. Kil, "Multilayer feedforward potential function network" IEEE International Conference on Neural Networks, vol.1, pp. 161-171, Jul 1988

[9] S. Haykin, *Neural Networks: A comprehensive Foundation*, MacMillan,1994

[10] T. Kailath, "The Divergence and Bhattacharyya Distance Measures in Signal Selection" IEEE Transactions on Communications, Vol. 15, Issue 1, Feb 1967, pp. 52-60.