Optimizing fusion architectures for limited training data sets

B. A. Baertlein and A. H. Gunatilaka

The Ohio State University ElectroScience Laboratory

1320 Kinnear Road, Columbus, OH 43212

ABSTRACT

A method is described to improve the performance of sensor fusion algorithms. Data sets available for training fusion algorithms are often smaller than desired, since the sensor suite used for data acquisition is always limited by the slowest, least reliable sensor. In addition, the fusion process expands the dimension of the data, which increases the requirement for training data. By using structural risk minimization, a technique of statistical learning theory, a classifier of optimal complexity can be obtained, leading to improved performance. A technique for jointly optimizing the local decision thresholds is also described for hard-decision fusion. The procedure is demonstrated for EMI, GPR and MWIR data acquired at the US Army mine lanes at Fort A.P. Hill, VA, Site 71A. It is shown that fusion of features, soft decisions, and hard decisions each yield improved performance with respect to the individual sensors. Fusion decreases the overall error rate (false alarms and missed detections) from roughly 20% for the best single sensor to roughly 10% for the best fused result.

Keywords: Sensor fusion, statistical learning theory, support vector machines, pattern classification

1. INTRODUCTION

Sensor fusion has been proposed as a means of meeting stressing requirements in detection of land mines. Fusion offers the potential for increased probability of detection, decreased false alarm rates, and operation in a broader range of environments. A very large number of sensor technologies have been proposed for mine detection,[1] and that number continues to grow.

Mine detection, whether done with a single sensor of a fused suite, can be viewed as a classification problem in which sensor data must be classified as representing mines or clutter. Suppose that we are given sensor data $\mathbf{x}$ with which to determine the truth of hypotheses H_k that describe the presence of absence of a mine. We do so by choosing H_k to maximize the Bayes' risk or, when only the probability of error is important, by choosing the hypothesis that maximizes the a posteriori probability $\Pr(H_k|\mathbf{x})$. A fundamental tenet of sensor fusion (or of any classification process) is that more information cannot degrade performance. It is easy to prove that additional independent sensor data $\mathbf{y}$ that is positively correlated with H_k (i.e., for which $\Pr(\mathbf{y}|H_k) \geq \Pr(\mathbf{y})$) will produce $\Pr(H_k|\mathbf{x},\mathbf{y}) \geq \Pr(H_k|\mathbf{x})$.

The probabilities appearing in the above expressions are unknown and must be estimated from data. All classification problems require training data from which the desired classification process is learned.* The quantity and quality of these training data have a strong influence on the performance of the classifier. Although many classifiers are known to produce optimal performance asymptotically (i.e., when the amount of training data is infinite), performance based on a finite training set may fail to meet expectations. The problem of limited training data is acute for mine detection in general, and for sensor fusion in particular. When multi-sensor data collections are performed, the data set useful for fusion is limited by the performance of the slowest or least reliable sensor.

The usefulness of a sensor-fused system (or any classifier) depends on its ability to generalize, i.e., to detect mines in data not seen previously. It is well known that problems with limited training data manifest themselves during generalization.[2] Classifiers that perform well on their training data but generalize poorly typically employ an architecture or an approximation to a decision surface that is too complex for the training data. Problems of this type arise in many classifiers, including k-nearest neighbor and neural networks.

Corresponding author: B.A.B. (614) 292-0076 (voice), (614) 292-7297 (fax), baertlein.1@osu.edu

*A sufficiently accurate physical model can supply the equivalent of training data, but for many mine detecting sensors the random environment has a strong influence on the sensor data, leading to an impractical number of model variables.

In this work we describe methods of optimizing the design of a classifier for a given training data set. For feature-level fusion we employ recent developments in statistical learning theory (SLT),[3] which permits us to bound the performance of classifiers designed with limited training data. Some essential aspects of SLT as it relates to mine detection are described in Section 2. The architecture of the fusion algorithm (classifier) determines its complexity, which can be bounded by SLT. The Vapnik-Chervonenkis (VC) dimension of the classifier provides a means of assessing the classifier complexity. For any classification problem, there is an optimum complexity. Unfortunately, it is largely impossible to estimate the VC dimension of most classifiers. One can implement SLT using a new form of classifier known as the support vector machine (SVM), for which the complexity is readily controlled. It has been observed that SVMs exhibit performance that meets or exceeds that of other classifiers. SVMs are described in Section 3. We describe in Section 4 a method of jointly optimizing the individual decision thresholds for decision-level fusion of hard decisions. In Section 5 we apply these methods to multi-sensor demining data collected at The US Army mine lanes at Fort A.P. Hill, VA (Site 71A). The data comprise samples of surrogates and buried mines with known positions. The sensor suite used included an EMI sensor, a GPR, and an MWIR camera. Concluding remarks appear in Section 6.

2. STATISTICAL LEARNING THEORY

Statistical learning theory, also referred to as Vapnik-Chervonenkis (VC) theory, has been under development since the 1970s. In this section we summarize the relevant parts of that theory. Descriptions of SLT have been given by Vapnik at an overview level[3] and at a deeper level.[4] Review works by Schölkopf et al.[5] and by Burges[6] may also be consulted for details not presented here.

Consider the following problem: Given N i.i.d. samples of training (sensor) data $\mathbf{x}_i$ with true classification y_i (e.g., $y_i = 1$ if sample $\mathbf{x}_i$ corrresponds to a mine and $y_i = 0$ otherwise), we wish to discover the classifier (function) $y = F(\mathbf{x})$ that will return the true identity y when presented with an input sample $\mathbf{x}$. We approximate this function by using training data to estimate parameters $\boldsymbol{\alpha}$ for a family of functions $F(\mathbf{x}; \boldsymbol{\alpha})$. The risk or expected loss for the classifier is given by

$$\mathcal{R}(\boldsymbol{\alpha}) = \int d\mathbf{x} \int dy L(y, F(\mathbf{x}; \boldsymbol{\alpha})) \Pr(\mathbf{x}, y) \tag{1}$$

where L is a loss function. The function $L(y, y')$ provides a measure of the "distance" between the true output y and the estimate $y' = F(\mathbf{x}, \boldsymbol{\alpha})$. For mine detection we are primarily interested in a loss function of the form

$$L(y, F(\mathbf{x}, \boldsymbol{\alpha}) = \begin{cases} C_{01} & y = 0,\ F(\mathbf{x}, \boldsymbol{\alpha}) = 1 \qquad \text{(false alarm)} \\ C_{10} & y = 1,\ F(\mathbf{x}, \boldsymbol{\alpha}) = 0 \qquad \text{(missed detection)} \\ 0 & y = F(\mathbf{x}, \boldsymbol{\alpha}), \qquad \text{(correct decision)} \end{cases} \tag{2}$$

where C_{01} and C_{10} are the costs of a false alarm and missed detection respectively. Clearly $C_{01} \ll C_{10}$, but it is not obvious how these costs should be assigned. In this work, we assume $C_{01} = C_{10} = 1$, in which case $\mathcal{R}$ is the probability of error. The extension to other cases is straightforward.

Since $\Pr(\mathbf{x}, y)$ in Eq. (1) is unknown, we are forced to estimate the true risk $\mathcal{R}$ from the available training data. We define the empirical risk

$$\mathcal{R}_{emp}(\boldsymbol{\alpha}) = \frac{1}{N} \sum_{n=1}^{N} L(y_i, F(\mathbf{x}; \boldsymbol{\alpha})) \tag{3}$$

Note that $\mathcal{R}_{emp}$ does not involve the probability density $\Pr(\mathbf{x}, y)$. Taking $C_{01} = C_{10} = 1$ makes $\mathcal{R}_{emp}$ an estimate of the classifier error rate.

Statistical learning theory addresses the relation between the true risk $\mathcal{R}$ and the empirical risk $\mathcal{R}_{emp}$. For a fixed classifier (fixed $\boldsymbol{\alpha}$) the empirical risk will always be less than the true risk, since one can develop a classifier that fits a finite set of training data arbitrarily well. The following bound can be derived

$$\mathcal{R}(\boldsymbol{\alpha}) \leq \mathcal{R}_{emp}(\boldsymbol{\alpha}) + \Phi(h/N) \tag{4}$$

where Φ is a confidence interval and h is the so-called "VC dimension." The confidence interval has the remarkable property that it is independent of the unknown probability distribution $\Pr(\mathbf{x}, y)$. The precise definition of h is somewhat technical, but it can be regarded as the number of training samples that can be correctly classified by

$F(\mathbf{x}, \boldsymbol{\alpha})$, i.e., the learning capacity of the classifier. A plot of these quantities appears in Figure 1. As the complexity of the classifier increases, the empirical risk (as measured by classifier performance on the training data) decreases. Simultaneously, there is less confidence (and a greater error rate) for more complex classifiers. In practice, we can use the resubstitution performance of the classifier to estimate $\mathcal{R}_{emp}$ and a validation set to estimate $\mathcal{R}$. In general, there exists an optimal value of h/N. Classifiers that have too little capacity are unable to learn the training data. Classifiers that are too complex will learn the training data well, but will have poor generalization capability. The objective of this work is to determine the level of classifier complexity that will minimize the true risk.

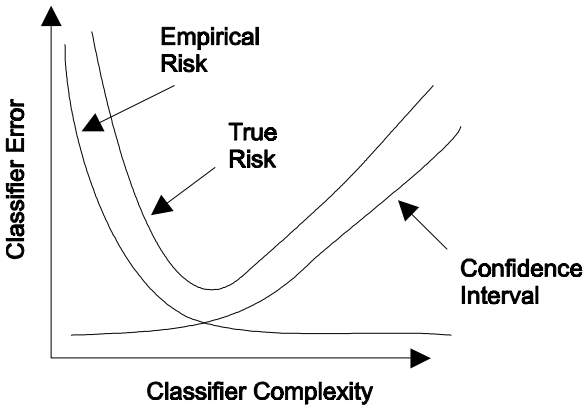


Figure 1. Components of classifier error.

The confidence interval $\Phi(h/N)$ is determined from the convergence rate of $\mathcal{R}_{emp} \rightarrow \mathcal{R}$ as $N \rightarrow \infty$. Previous work has produced a variety of expressions for this bound. For a loss function L bounded above by unity, Vapnik[4] has shown that with confidence η the value of Φ is given by

$$\Phi = \frac{\epsilon^2}{2}\left(1 + \sqrt{1 + \frac{\mathcal{R}_{emp}}{\epsilon^2}}\right) \tag{5}$$

where ϵ is given by

$$\epsilon = 2\sqrt{\frac{h}{N}\left[\ln\left(\frac{2N}{h}\right) + 1\right] - \frac{1}{N}\ln\eta} \tag{6}$$

For small empirical risk $\mathcal{R}_{emp}$ we have $\Phi \approx \epsilon^2$ while for large empirical risk we have $\Phi \sim \epsilon/2$. This last approximation is often cited in the literature, because of its simple form. It predicts that the confidence interval Φ decreases with increasing training data N and with decreasing classifier capacity h. Note also that Φ is an upper (worst case) bound. It may well be possible to design a classifier whose performance beats this bound.

Finding the optimum classifier consists in balancing the empirical risk (which decreases with increasing classifier capacity) and the confidence interval (which increases with increasing classifier capacity). There exist a variety of methods for designing classifiers, but in general optimizing classifier performance is a trial and error process. Although one is guaranteed that an optimum classifier exists (since the risk is bounded below), there is little guidance on how it can be constructed.

The concept illustrated in Figure 1 suggests a principled method for designing a classifier with optimal performance on a given training data set. If one can develop a family of classifiers $F_1, F_2, \ldots$ with increasing VC dimension $h_1 < h_2 < \ldots$, then by testing these classifiers we can determine the value of h that minimizes the true risk, i.e., we can identify the minimum in the true risk curve of Figure 1. This procedure is known as structural risk minimization (SRM),[7] and it is used most effectively in concert with support vector machines, described below.

3. SUPPORT VECTOR MACHINES

The existence of an optimal value of VC dimension h is of considerable theoretical interest, but it is of limited practical value since h is essentially impossible to estimate for any but the simplest classifiers. One classifier for which h can be determined is a linear hyperplane classifier

$$F(\mathbf{x}, \mathbf{w}, b) = \text{sgn}\,(\mathbf{x} \cdot \mathbf{w} + b) \tag{7}$$

where $\mathbf{w}$ is a vector that defines the normal to the hyperplane and b is a bias term.

Support vector machines (SVMs) are based on the concept of the optimal separating hyperplane, in which the value of $\mathbf{w}$ is selected to maximize class separation (also known as "margin"). This is achieved by minimizing $||\mathbf{w}||$ while simultaneously requiring that correct decisions are produced for the training data. We require

$$\text{minimize} \quad ||\mathbf{w}||^2 \qquad \text{subject to} \quad y_i(\mathbf{x}_i \cdot \mathbf{w} + b) \geq 1, \qquad i = 1, 2, \ldots, N \tag{8}$$

A Lagrangian formulation for the problem leads to

$$L(\mathbf{w}, b) = \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{N} \alpha_i [y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1] \tag{9}$$

where the α_i are Lagrange multipliers. This quantity must be minimized with respect to $\mathbf{w}$ and b, and maximized with respect to the α_i. The minimization requirements impose linear constraints as follows:

$$\frac{\partial}{\partial b} L = 0 \qquad \Rightarrow \sum_{i=1}^{N} \alpha_i y_i = 0 \tag{10}$$

$$\frac{\partial}{\partial w_j} L = 0 \qquad \Rightarrow \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i = \mathbf{w} \tag{11}$$

It can be shown that for a linearly separable problem, all but a few of the α_i will be identically zero. The training data $\mathbf{x}_i$ corresponding to these nonzero weights are known as the support vectors. They define the hyperplane, and the remaining training data are superfluous to the classifier. Substituting equations (10) and (11) into the Lagrangian allows us to eliminate $\mathbf{w}$ in favor of the α_i, leading to the dual form of Eq. (8). Using a vector notation for the α_i we obtain the following quadratic programming problem:

$$\text{maximize} \quad W(\boldsymbol{\alpha}) = \mathbf{e}^T\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} \qquad \text{subject to} \quad \mathbf{y}^T \cdot \boldsymbol{\alpha} = 0; \qquad \alpha_i \geq 0 \tag{12}$$

where $\mathbf{e} = [1\ 1\ \cdots\ 1]^T$ and

$$[\mathbf{Q}]_{ij} = y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \tag{13}$$

Numerical optimzation is used to determine $\boldsymbol{\alpha}$, and the decision function is

$$F(\mathbf{x}; \boldsymbol{\alpha}, b) = \text{sgn}\left(\sum_{i=1}^{N} [y_i \alpha_i \mathbf{x} \cdot \mathbf{x}_i + b]\right) \tag{14}$$

When the classes overlap and the data are not separable by a hyperplane, one can show that the appropriate formulation is identical to that presented above if we introduce "slack" variables $\zeta_i \geq 0$ such that the problem becomes

$$\text{minimize} \quad ||\mathbf{w}||^2 + C\sum_{i=1}^{N} \zeta_i \qquad \text{subject to} \quad y_i(\mathbf{x}_i \cdot \mathbf{w} + b) \geq 1 - \zeta_i, \qquad i = 1, 2, \ldots, N \tag{15}$$

where $C > 0$ is a user-defined constant. Taking $C \to \infty$ produces the separable case, while $C \to 0$ reduces the penalty for class overlap. The minimization proceeds as described above for the separable case, except that the α_i are now constrained as

$$0 \leq \alpha_i < C \tag{16}$$

It is instructive to compare SVMs with Fisher's linear discriminant,[8] another common hyperplane classifier. The normal to the Fisher hyperplane $\mathbf{w}_f$ for a two-class problem having means ($\mathbf{m}_0$,$\mathbf{m}_1$) and covariance matrices ($\mathbf{C}_0$,$\mathbf{C}_1$) satisfies the generalized eigenvector problem[9]

$$[\mathbf{C}_B - \lambda(\mathbf{C}_1 + \mathbf{C}_1)]\mathbf{w}_f = 0 \tag{17}$$

where $\mathbf{C}_B$ is the between-class scatter matrix

$$\mathbf{C}_B = \frac{N_0 N_1}{N}[\mathbf{m}_0 - \mathbf{m}_1][\mathbf{m}_0 - \mathbf{m}_1]^T \tag{18}$$

Since $\mathbf{C}_B$ has rank one, this equation has one non-zero eigenvector, and we find

$$\mathbf{w}_f = (\mathbf{C}_1 + \mathbf{C}_1)^{-1}[\mathbf{m}_0 - \mathbf{m}_1] \tag{19}$$

which is parallel to the path between the mean vectors. Although the Fisher discriminant is known to be optimal for Gaussian distributions with equal covariances, it does not necessarily produce an optimal separation for a finite data set. In general, the SVM hyperplane is not parallel to the Fisher hyperplane.

Linear classifiers are seldom optimal in practice, and higher-order approximations to the decision surface are commonly employed. The hyperplanes used in SVMs can also be extended to non-planar surfaces. To do so, the input data $\mathbf{x}_i$ are projected into higher dimensions by using a nonlinear tranformation $\Psi(\mathbf{x})$. In these high dimensional spaces, the data are more likely to be separable by hyperplanes. The problem formulation parallels that given above with the substitution $\mathbf{x} \rightarrow \Psi(\mathbf{x})$.

A classical problem with higher-order classifiers is dimensionality. Consider a simple polynomial transformation of the form $\Psi(\mathbf{x}) = \{1, x_1, x_2, ...x_m, x_1^2, x_1x_2, x_1x_3, ...\}$. The product terms are each inputs to the classifier. It is evident that the resulting input set has high dimensions, for which the memory, computation, and training data requirements are large. An ingenious method is used to avoid this for SVMs. Note that the SVM requires only inner products of the projected data of the form $\Psi(\mathbf{x}) \cdot \Psi(\mathbf{y})$. Mercer's theorem[10] implies that (with certain mild restrictions) such inner products can be written in terms of a symmetric kernel function $k(\mathbf{x}, \mathbf{y})$. Hence, given a transformation Ψ, we can replace the inner product $\Psi(\mathbf{x}_i) \cdot \Psi(\mathbf{x}_j) \rightarrow k(\mathbf{x}_i, \mathbf{x}_j)$. Since appropriate transformations Ψ are seldom evident, it is attractive to use Mercer's theorem in "reverse", i.e., to assume a convenient form for $k(\mathbf{x}_i, \mathbf{x}_j)$ without regard for the implied function Ψ. Thus, the above-described formulation still applies if the matrix $\mathbf{Q}$ becomes

$$[\mathbf{Q}]_{ij} = y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \tag{20}$$

Functions used for $k(\cdot, \cdot)$ include polynomials, radial basis functions, splines, and others.

4. OPTIMIZING FUSION OF HARD DECISIONS

In general, optimal fusion of hard decisions is challenging. Each sensor produces a binary decision u_i on the basis of a sensor-specific threshold t_i of a test statistic (e.g., the log-likelihood ratio). For an optimal fusion strategy one must simultaneously define a relation among the decision thresholds (t_1, t_2, t_3) (i.e., the local decision strategies) and a fusion rule $\Pr(H_k|u_1, u_2, u_3)$ (the global decision strategy). In general, numerical optimization is required,[11] although some simple expressions are available if one assumes that the sensors are independent. In a previous fusion study, we found the performance of the latter approach to be poor,[12] and in this work we explored alternative methods.

A straightforward (but not necessarily optimal) local decision strategy is to remap all of the individual sensor test statistics to a common range and then use a common threshod $t \equiv t_1 \equiv t_2 \equiv t_3$. Although this technique can work well in some situations, it is adversely affected by outliers in the data or drastically different performance from sensor to sensor.

A somewhat better approach is to define a monotonic function $G(t; \boldsymbol{\beta}_m)$ with parameters $\boldsymbol{\beta}_m$ to remap the test statistics for each sensor m. A common threshold τ is used for the remapped thresholds $G(t; \boldsymbol{\beta}_m)$ for all sensors. Since (by assumption) the data sets are small, numerical optimization of the $\boldsymbol{\beta}_m$ is practical. In addition, since the number of sensors involved is also small, one can simultaneously conduct an exhaustive search over all possible fusion rules $\Pr(H_k|u_1, u_2, u_3)$. The SRM technique is not used explicitly in this design of the hard decision fusion algorithm, but the classifiers for the individual sensors may be optimized using SRM.

5. RESULTS

We used structural risk minimization to optimize the performance of mine detectors based on both single sensors and a fused sensor suite. The threshold-remapping methods were also used to optimize hard decision fusion. In this section we describe the data set, review some sensor-specific processing, and present the results.

During July 1999 we acquired multisensor data at Fort A.P. Hill, VA. Data were collected at the calibration lanes of Site 71A, which comprises a 25 m by 5 m area with mines or clutter objects buried at the center of grid cells having dimensions 1 m by 1 m. Some important features of this test site are (1) extraneous metal has been largely removed from the area, which reduces EMI clutter; (2) the location of the mine (or clutter) item can be accurately located, which obviates problems with sensor positioning; and (3) although real mines have been emplaced, in most cases the explosive was removed from the mines, which affects its thermal signature.

We acquired data over a portion of the site that comprised 27 deactivated mines and 32 clutter objects. The sensors used included a Schiebel AN 19/2, an OSU-developed ground penetrating radar,[13] a MWIR sensor, and a LWIR sensor. In the results that follow, we have replaced the Schiebel data with data from the GEM-3 sensor,[14] collected by Duke University.[15] During most of our data collection, conditions were not conducive to IR data collection. The weather was overcast and rain fell occasionally. Nonetheless, we did manage to acquire some useful MWIR data during one night-time collection. In the discussion that follows, we do not consider the LWIR data.

5.1. Supporting Processing

The data collected by each sensor have a different format and require different processing to suppress clutter and to extract features. In this section we briefly describe that processing.

5.1.1. GEM-3 Sensor

The GEM-3 data comprise samples taken at ten points spaced 2 inches apart in a "+" pattern over each putative target, with 5 samples taken in a left-right path and 5 samples taken in a fore-aft path. At each point, in-phase and quadrature magnetic field measurements were acquired at 20 frequencies logarithmically spaced from 270 Hz to 23.79 kHz. For the 7/7/98 data set used here, data were acquired at 44 locations in the Site 71A calibration area, where the identify of the targets are known. Background data, required to correct for sensor baseline drift, were acquired between measurements. An extensive discussion of the GEM-3 data has been presented by Gao et al.,[16] which can be consulted for details not discussed here.

Processing of the GEM-3 data included correcting for the sensor background and converting the sensor output to a quantity proportional to the complex field amplitude. The resulting data set (200 complex values at each target) was reduced by computing the energy at each spatial position, leading to ten values for each target. An example EMI data vector is shown in Figure 2. For perfectly centered, symmetric targets these signatures should display symmetry about samples 3 and 8 (i.e., about the center of the left-right and fore-aft scans) and samples 1-5 (the left-right scan) and 6-10 (the fore-aft scan) should be identical.

5.1.2. GPR Sensor

The OSU GPR is a down-looking sensor using a novel dielectric rod antenna.[13] The antenna was mounted to a linear positioner, which was scanned over the target location. Data were acquired at 1-2 cm sample intervals over each target cell. Linear scans were acquired over 59 mines and clutter sites. The system uses a wide bandwidth (1-6 GHz), and after pulse compression data similar to Figure 3(a) are observed. The strong near-horizontal band in this image is the ground reflection. Using a recently developed technique,[17] that reflection can be significantly reduced, leading to the data shown in Figure 3(b). In the latter result, the characteristic hyperbolic arcs generated by the target are evident. The data used in processing comprised a subsampled version of the time-domain output acquired directly through the center of each signature.

5.1.3. MWIR Sensor

The MWIR sensor used in this collection was a COTS camera (Cincinnatti Electronics, IRRIS 160ST) operating in the spectral band 2.2-4.6 μm. The instrument's focal plane comprises 160×120 pixels, with an NEΔT of 0.025K. To avoid clutter produced by reflected sunlight, the sensor was operated at night. During data acquisitioin the sensor was positioned at a fixed height and distance from each putative target site. The camera was aimed at a ground location a known distance from the camera's ground-projected center. Using the (known) field of view, one can determine

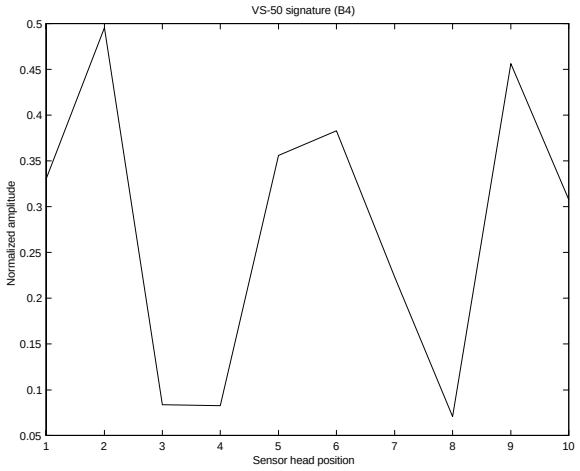
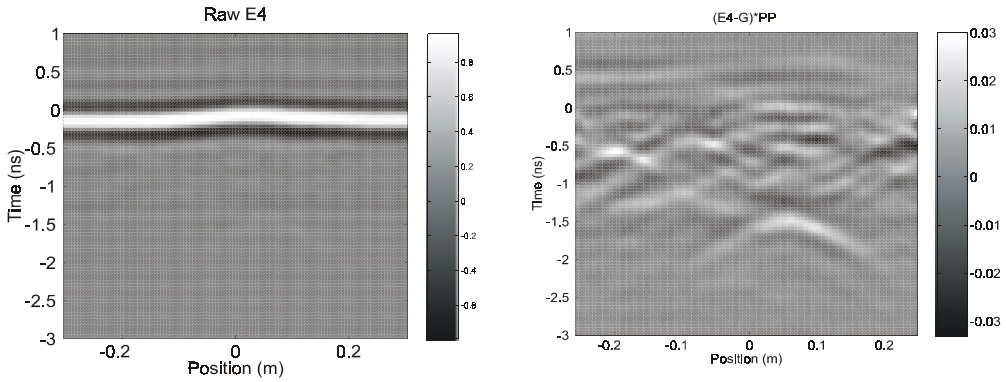


Figure 2. Example GEM-3 data.



(a) Raw GPR image (b) Clutter-reduced image

Figure 3. Comparison of raw and clutter-reduced data generated from measurements over a three-inch deep VS-50 mine.

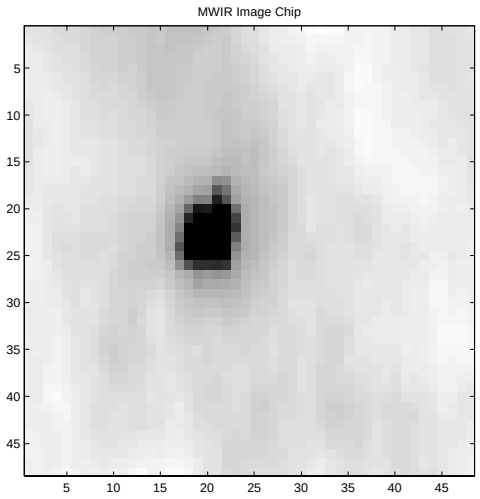


Figure 4. Example MWIR data. A M14 mine flush buried is shown.

the position of each pixel on the ground. After remapping the image to eliminate the effects of perspective, data similar to Figure 4 are obtained. Features were extracted from these data using a model-based technique described previously.[18] IR clutter is normally distributed, and a maximum likelihood (nonlinear least squares) technique was used to estimate model parameters used as features.

5.2. Individual Sensor Performance

A review of the data acquired by the GEM-3, GPR, and MWIR camera revealed that data from all three sensors were available at 42 grid cells, which comprised equal number of target and clutter samples. Features were extracted from the data acquired by each sensors as described above.

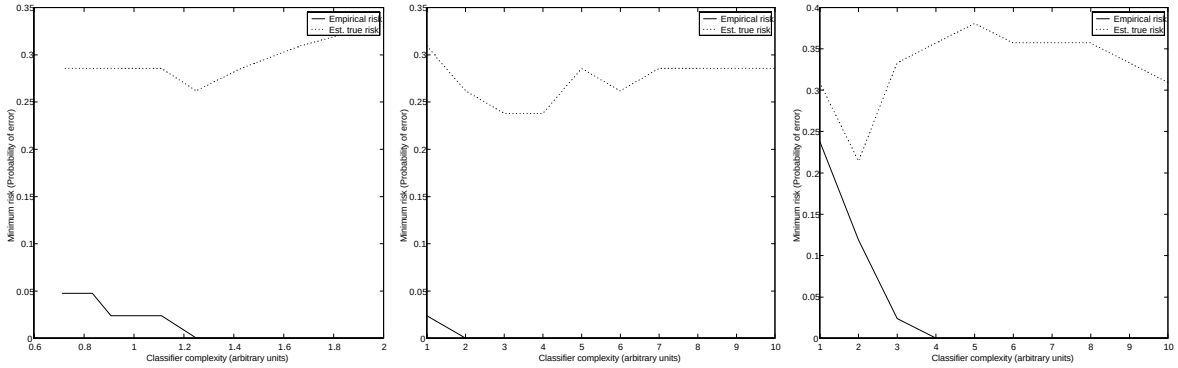
The SRM method was used to identify an optimum classifier complexity for each sensor. For this small data set we used the "leave-one-out" method (a form of cross-validation, or resampling) to estimate the true risk for a range of classifier complexity. The resubsitution method was used to determine the classifier empirical risk. Both polynomial classifiers and radial basis function classifiers[†] were examined, and the design producing the best performance was used. In general, the radial basis function classifiers had a small advantage, but the performance of other types of nonlinearities produced comparable results. This finding is similar to that reported in the literature for other SVM applications. In our tests we found that the slack variable weight C had a minimal effect on classifier performance. We used $C = 10$ for all tests. The risk calculation for each sensor is shown in Figure 5. In each case the empirical risk decreases with increasing complexity, and a minimum occurs in the estimated true risk (cf. Figure 1). The minimum risk (equal to the error rate) for the best sensor (MWIR) is slightly larger than 20%.

Using the classifier design that produced the minimum true risk, we computed the ROC curves as shown in Figure 6. We observe that no sensor has an overwhelming advantage in detection.

5.3. Fusion Performance

We explored fusion of sensor features, soft decisions (classifier outputs produced for individual sensors), and hard decisions (the result of thresholding individual sensor classifier outputs). Feature-level fusion is a straightforward process. We concatenated the feature vectors for all sensors and trained a SVM classifier. Fusion of soft decisions can be performed using a number of techniques.[12] We opted to form a hierarchical classifier in which the outputs of the single sensor classifiers are supplied to another SVM.

[†]Although the VC dimension of a radial basis function classifier is infinite, the width of the basis functions provides a degree of control over its complexity.[6]



(a) GEM-3 risk (b) GPR risk (c) MWIR risk

Figure 5. Risk calculations for the individual sensors.

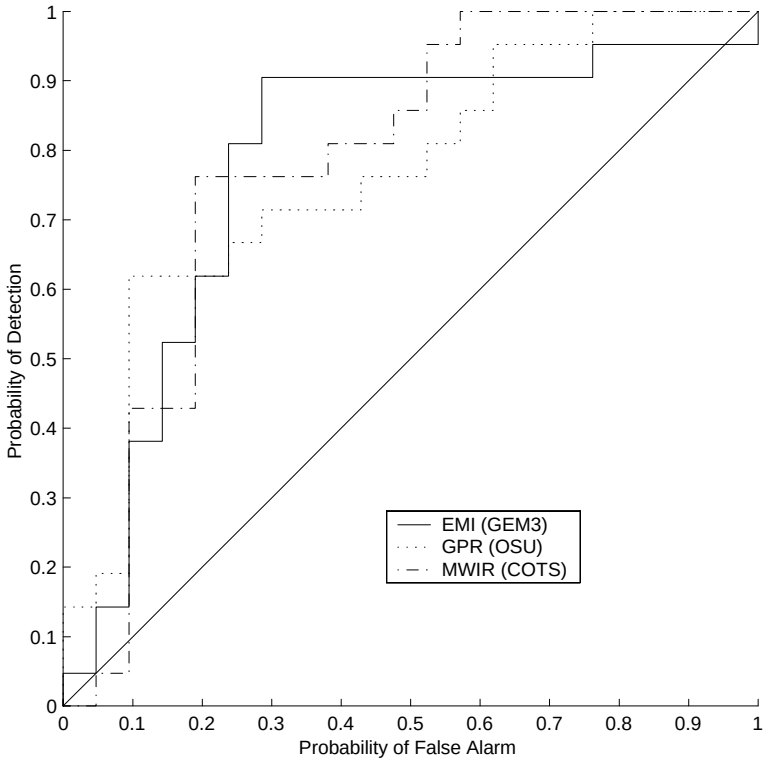
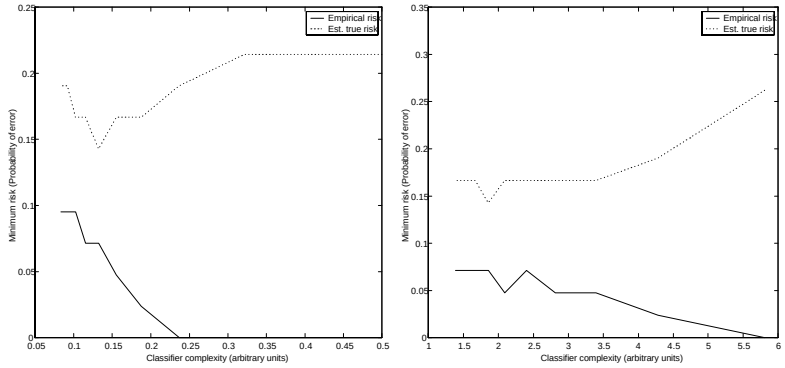
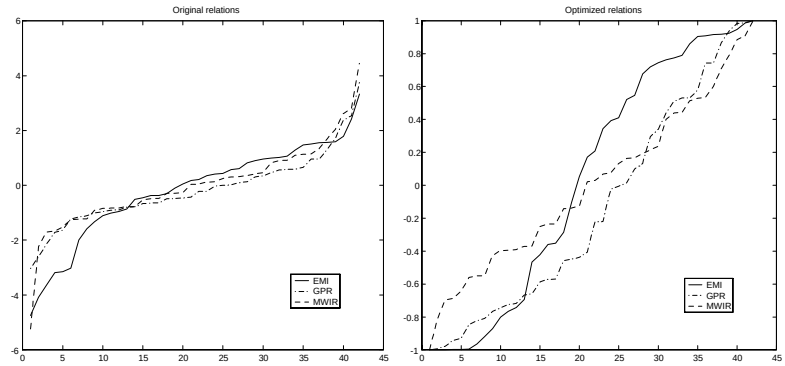


Figure 6. ROC curves for the individual sensors.



(a) Feature-level fusion risk (b) Soft-decision fusion risk

Figure 7. Risk calculations for the fused sensor suite.



(a) Original test statistics (b) Test statistics after remapping

Figure 8. Test statistics (a) before and (b) after the optimization process.

The SRM approach was used to optimize both the feature-level and soft decision-level fusion classifiers. The resulting risk data are shown in Figure 7. Again, we note the presence of minima, which denote the optimum classifier complexity for this data set. The minimum risk is approximately 15% for these classifiers.

The SVM-optimized results shown in Figure 6 were the basis for optimizing hard decision fusion. After sorting the classifier outputs into ascending order, the test statistics for these sensors trace out the curves given in Figure 8. That figure also includes the remapped test statistics after the optimization process described in Section 4. The remapping functions G were the product of a logistic function (followed by scaling to a common range) and Chebyshev polynomials with numerically optimized weights $\boldsymbol{\beta}_m$. The optimization process tends to increase the slope of these curves near the decision boundary (the zero crossing), which tends to increase the decision margin. Because the performance of these sensors are comparable, a majority voting technique was used for $\Pr(H_k|u_1, u_2, u_3)$. That rule is near optimal for more than two sensors of comparable performance. The true risk estimated for this classifier is 9.5%. It is interesting to note that for this data set, a simple linear remapping of the raw test statistics does nearly as well (12% error).

ROC curves are shown in Figure 9 for feature-level, soft-decision level, and hard-decision level fusion. We observe that feature-level and soft decision fusion produce comparable performance. In contrast to the findings of our previous

work,[12] hard-decision fusion performs better than either feature-level fusion or hard-decision fusion. We attribute this improvement to (1) our decision to reject the independent sensor hypothesis and (2) the fact that no one sensor has an overwhelming performance advantage (which tends to reduce the benefit of fusion under the independent sensor hypothesis).[11] These findings bear further investigation. The benefit of fusion can be quantified by comparing the risk for the best fused suite (hard decision fusion at roughly 10%) and the best individual sensor (MWIR at roughly 20%).

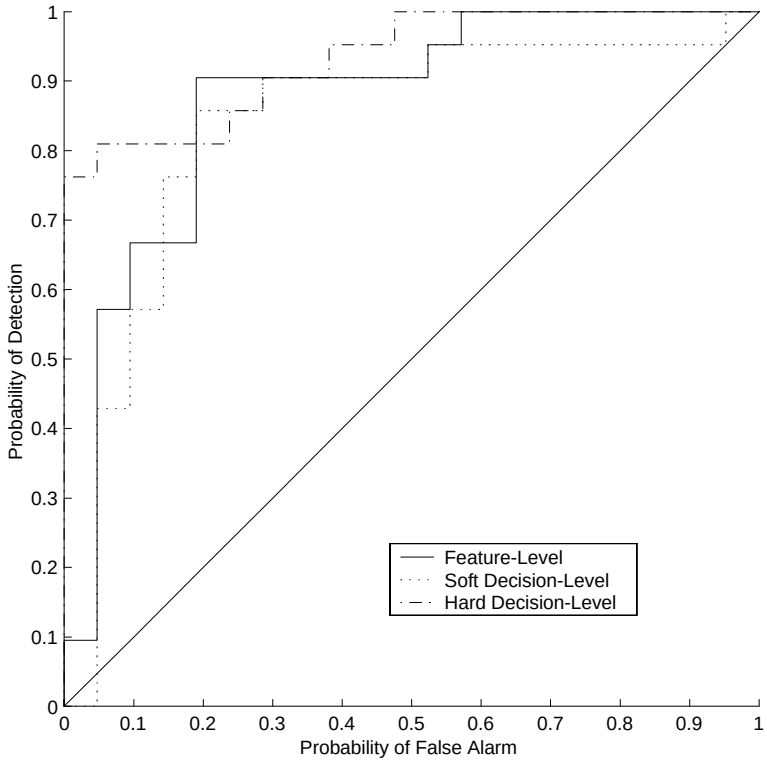


Figure 9. ROC curves for three forms of fusion.

6. CONCLUDING REMARKS

Data sets available for the design of sensor-fused mine detection algorithms are typically small, which has an adverse effect on fusion algorithm design and detection performance. We have described an approach based on structural risk minimization that permits us to extract the best possible fusion performance from these small data sets.

Tests of the algorithm described here for 42 samples of EMI, GPR and MWIR data suggest that fusion produces a measureable benefit in performance. We found that the net error rate (missed detections and false alarms) decreased by roughly a factor of two when the best individual sensor is compared to the best fused result.

ACKNOWLEDGMENTS

The authors would like to thank the staff of E-OIR Measurements, Inc. and the Joint UXO Coordinating Office for their support during data collections at Fort A. P. Hill, Site 71A. We would also like to thank L. Collins at Duke University for her assistance in obtaining and interpreting the GEM-3 data.

This project was supported by funds from Duke University under an award from the ARO (the OSD MURI program). The findings, opinions and recommendations expressed therein are those of the authors and are not necessarily those of Duke University or the ARO.

REFERENCES

1. C. Stewart, "Summary of mine detection research (U)," technical report 1636-TR, U.S. Army Engineer Research and Development Laboratories, Corps of Engineers, Fort Belvoir, VA, May 1960. Volume 1, DTIC AD320124.
2. R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley, New York, NY, 1973.
3. V. Vapnik, *The Nature of the Statistical Learning Theory*, Springer Verlag, New York, NY, second ed., 2000.
4. V. Vapnik, *Statistical Learning Theory*, Wiley, New York, NY, 1998.
5. B. Schölkopf, C. J. C. Burges, and A. J. Smola, *Advances in Kernel Methods: Support Vector Learning*, MIT Press, Cambridge, MA, 1999.
6. C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery* **2**(2), pp. 1–49, 1998.
7. V. Vapnik, *Estimation of Dependences Based on Empirical Data (in Russian)*, Nauka, Moscow, 1979. English translation: Springer, New York, 1982. Also see Ch. 4, Vapnik, 2000.
8. R. A. Fisher, "The use of multiple measures in taxonomic problems," *Ann. Eugenic* **7**, pp. 179–188, 1936.
9. J. J. W. Sammon, "An optimal discrimination plane," *IEEE Trans. Computers* , pp. 826–829, Sept. 1970.
10. J. Mercer, "Functions of positive and negative type and their connection with the theory of integral equations," *Philos. Transactions of the Royal Society of London* **A209**, pp. 415–446, 1909.
11. P. K. Varshney, *Distributed Detection and Data Fusion*, Springer-Verlag, New York, NY, 1997.
12. A. H. Gunatilaka and B. A. Baertlein, "Comparison of predection and postdetection fusion for mine detection," in *Detection and Remediation Technologies for Mines and Minelike Targets IV*, A. C. Dubey, J. F. Harvey, J. T. Broach, and R. E. Dugan, eds., *SPIE* **3710**, pp. 1212–1223, 1999.
13. S. Nag, L. Peters, I. J. Gupta, and C.-C. Chen, "Ramp response for the detection of anti-personnel mines," in *Detection and Remediation Technologies for Mines and Minelike Targets IV*, A. C. Dubey, J. F. Harvey, J. T. Broach, and R. E. Dugan, eds., *SPIE* **3710**, pp. 1313–1322, 1999.
14. I. J. Won, D. A. Keiswetter, and D. R. Hansen, "A monostatic broadband electromagnetic induction sensor," *Journal of Environmental and Engineering Geophysics* **2**, pp. 53–64, Aug. 1997.
15. L. M. Collins, "Statistical signal processing for demining: Experimental validation," tech. rep., Duke University, 1998.
16. P. Gao, L. Collins, J. Moulton, L. Makowsky, R. Weaver, D. Keiswetter, and I. J. Won, "Enhanced detection of landmines using broadband EMI data," in *Detection and Remediation Technologies for Mines and Minelike Targets V*, A. C. Dubey, J. F. Harvey, J. T. Broach, and R. E. Dugan, eds., *SPIE* **3710**, pp. 2–13, April 5-9 1999.
17. A. H. Gunatilaka and B. A. Baertlein, "A subspace decomposition technique to improve GPR imaging of anti-personnel mines," in *Detection and Remediation Technologies for Mines and Minelike Targets V*, A. C. Dubey, J. F. Harvey, J. T. Broach, and R. E. Dugan, eds., *SPIE* **4038-111**, April 24-28 2000.
18. I. K. Sendur and B. A. Baertlein, "Techniques for improving buried mine detection in thermal IR imagery," in *Detection and Remediation Technologies for Mines and Minelike Targets IV*, A. C. Dubey, J. F. Harvey, J. T. Broach, and R. E. Dugan, eds., *SPIE* **3710**, pp. 1272–1283, 1999.