F3-C,E: Machine Learning & Sensor Management for High Throughput Screening

Abstract— This project is investigating the development of automated explosive detection and classification algorithms for high throughput screening. This is critical both in portal systems, where high throughput requires significant automated decision support, and in stand-off systems where the proliferation of multimodal data can overwhelm human interpretation. The project’s fundamental assumption is that it is too slow or costly to collect full sensor data on every object of interest, either for training, or during real-time operation. As a consequence, there are several important problems to address. In training, one needs to select which data will be used to train the decision algorithms in order to achieve robust performance. This is a problem known as active learning. In the real-time phase, one needs to use a hierarchy of sensing and classification strategies, based on relatively inexpensive early warning sensors, and adaptively select subsequent sensor measurements in order to arrive rapidly at an accurate classification decision. The long-range impact of this research will be the development of adaptive, high throughput screening algorithms for different combinations of sensing modalities that exhibit improved sensitivity/specificity.

I. PARTICIPANTS

<table>
<thead>
<tr>
<th>Faculty/Staff</th>
<th>Name</th>
<th>Title</th>
<th>Institution</th>
<th>Email</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>David Castañón</td>
<td>Professor</td>
<td>Boston University</td>
<td><a href="mailto:dac@bu.edu">dac@bu.edu</a></td>
<td>617.353.9880</td>
</tr>
<tr>
<td></td>
<td>Venkatesh Saligrama</td>
<td>Professor</td>
<td>Boston University</td>
<td><a href="mailto:srv@bu.edu">srv@bu.edu</a></td>
<td>617.353.1040</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Students</th>
<th>Name</th>
<th>Degree Pursued</th>
<th>Institution</th>
<th>Email</th>
<th>Intended Year of Graduation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Kirill Trapeznikov</td>
<td>PhD</td>
<td>Boston University</td>
<td><a href="mailto:k.trapeznikov@gmail.com">k.trapeznikov@gmail.com</a></td>
<td>2012</td>
</tr>
<tr>
<td></td>
<td>D. Motamed-Vaziri</td>
<td>PhD</td>
<td>Boston University</td>
<td><a href="mailto:deli@bu.edu">deli@bu.edu</a></td>
<td>2013</td>
</tr>
<tr>
<td></td>
<td>Joe Wang</td>
<td>PhD</td>
<td>Boston University</td>
<td><a href="mailto:joewang@bu.edu">joewang@bu.edu</a></td>
<td>2013</td>
</tr>
</tbody>
</table>

II. PROJECT OVERVIEW AND SIGNIFICANCE

In this project we will address the problem of cost-sensitive sequential learning, which is targeted at improving automated algorithm performance. The need for real-time throughput requires sequential and hierarchical inspection strategies where time-consuming inspections are used only for the most difficult cases. However, most research in machine learning focuses on the design of individual classification systems rather than a sequential or hierarchical network of classifiers. We are interested in the development of design theories for networks of classifiers that are throughput limited, to achieve superior classification performance.

There are many technical hurdles that need to be overcome in this research. First, many empirical multimodal data sets are unbalanced, containing fewer samples of explosive threats than non-explosive threats. Second, there are varying costs associated with different types of errors – the cost of a false alarm is not equal...
to the cost of a missed detection, and these misclassification costs can change over time. Third, there can be a significant amount of training data that is not labeled. This may happen, for instance, by collecting data from multi-sensor portals without a subsequent detailed examination of each individual item. Fourth, most classification theories in machine learning train classifiers to make terminal decisions, rather than to be part of sequential, adaptive decision processes.

A. Long-Range Impact

The long-range impact of this research will be the development of adaptive, high throughput screening algorithms for different combinations of sensing modalities that exhibit improved sensitivity and specificity.

In many applications including homeland security and medical diagnosis, decision systems are composed of an ordered sequence of stages. Each stage is associated with a sensor or a physical sensing modality. Typically, a less informative sensor is cheap (or fast) while a more informative sensor is either expensive or requires more time to acquire a measurement. In practice, a measurement budget (or throughput constraint) does not allow all the modalities to be used simultaneously in making decisions. The goal in these scenarios is to attempt to classify examples with low cost sensors and limit the number of examples for which a more expensive or time consuming informative sensor is required.

As an example consider the problem of explosives detection. In the first stage, an infra-red (IR) imager is used. The second stage is a more expensive and time consuming active millimeter wave (AMMW) scanner. The final third stage is a time consuming human inspection. In medical applications, first stages are typically non-invasive procedures (such as a physical exam) followed by more expensive tests (blood test, CT scan etc) and the final stages are invasive (surgical) procedures. Many such examples share a common structure (see Fig. 1) and we list some of its salient aspects below:

Sensors & Ordered Stages: Each stage is associated with a new sensor measurement or sensing modality. Multiple stages are an ordered sequence of sensors or sensor modalities with later stages corresponding to expensive or time-consuming measurements. In many situations there is often some flexibility in choosing a sensing modality from a collection of possible modalities. In these cases the optimal choice of sensing actions also becomes an issue. While our methodology can be modified to account for this more general setting, we considered a fixed order of stages and sensing modalities in this paper. This is justified on account of the fact that many of the situations we have come across consist of a handful of sensors or sensing modalities. Consequently, for these situations, the problem of choosing sensor ordering is not considered since one could by brute force enumerate and optimize over the different possibilities.

Reject Classifiers: Our sequential decision rules either attempt to fully classify an instance at each stage or “reject” the instance on to the next stage for more measurements in case of ambiguity. For example, in explosives detection, a decision rule, in the first stage, based on an IR scan, would attempt to detect whether or not...
a person is a threat and identify the explosive type/location in case of a threat. If the person is identified as a threat at the first stage it is unnecessary and potentially dangerous (the explosive could be detonated) to seek more information. Similarly in medical diagnosis if a disease is diagnosed at an early stage, it may be better to start early treatment rather than waiting for more conclusive tests.

In general multiple sensors can acquire different features on the same object. Each sensor has an acquisition cost (time, exposure, etc). In general less informative features are cheap (or fast), while more informative features are expensive or take time to acquire. We have a measurement budget (or throughput constraint) which precludes all the features to be used simultaneously

Our approach is to collect low cost features on an object, and identify ambiguous cases where additional sensors can help. We then classify easy cases with low cost features, and the harder cases with more expensive features.

Figure 2 illustrates some of the advantages of our scheme (adaptive) for a simple two stage example comparing our approach with an alternative scheme that first acquires measurements from all the sensors or sensing modalities, which we refer to as the centralized classifier. Our reject classifier observes the value of the first sensor, and determines to acquire the second sensor information if the value is in the green region. The resulting performance is equivalent to that of a classifier that used both sensors on all the samples, but utilizes the second stage sensor only for a fraction of the samples.

III. RESEARCH AND EDUCATION ACTIVITY

A. State-of-the-Art

Cost-Sensitive Learning:

Cost-Sensitive Learning has been studied in the Machine Learning community as early as MacKay (1992a). Our work is closely related to the so called prediction time active feature acquisition (AFA) approach in the area of cost-sensitive learning. The goal there is to make sequential decisions of whether or not to acquire a new feature to improve prediction accuracy. A natural approach is to formalize a problem as an Markov Decision Problem (MDP). Ji and Carin (2007); Kapoor and Horvitz (2009) model the decision process and infer feature dependencies while taking acquisition costs into account. Sheng and Ling (2006); Bilgic and Getoor (2007); Zubek and Dietterich (2002) study strategies for optimizing decision trees while minimizing acquisition costs. The construction is usually based on some purity metric such as entropy. Kanani and Melville (2008) propose a method that acquires an attribute if it increases an expected utility. However, all these
methods require estimating a probability likelihood that a certain feature value occurs given the features collected so far. While surrogates based on classifiers or regressors can be employed to estimate likelihoods, this approach requires discrete, binary or quantized attributes.

In contrast, our problem domain deals with high dimensional measurements (images consisting of million of pixels) and so we develop a discriminative learning approach and formulate a multi-stage empirical risk optimization problem to reduces measurement costs and misclassification errors. At each stage we solve the reject classification problem by factorizing the cost function into classification and rejection decisions. We then embed the rejection decision into a binary classification problem.

B. **Major Contributions**

Our contributions to date have been along several directions. First, we have developed new approaches for adaptive classifiers that can separate test cases into explosive threats, non-threats and ambiguous objects needing further attention. Classification is done in a Bayes optimal manner taking into account the relative costs of different types of errors plus the cost of additional tests. These techniques have been applied successfully for optical cancer biopsies using hyperspectral methods, increasing the sensitivity and specificity of the automated decisions. The underlying assumption is that there is an expensive, but very accurate, mode of detection (e.g. manual inspection) that should be invoked infrequently.

In our earlier work, we developed an extension of Support Vector Machine (SVM) classifiers for classifying samples belonging to one of two classes (threats vs. non-threats), where the classifiers include regions for delayed decisions in addition to regions for immediate classification. We introduced a new convex optimization formulation based on empirical risks for the design of these classifiers, and developed a class of coordinate ascent training algorithms to solve the design problem based on extensions of the Sequential Minimization algorithm (SMO). The formulation includes the cost of different error types as well as the costs of delaying decision to collect additional information.

Our recent work (Trapeznikov et. al, [ACML 2012, SSP2012]) is based on the so called Prediction Time Cost Reduction approach Kanani and Melville (2008). Specifically, we assume a set of training examples in which measurements from all the sensors or sensing modalities as well as the ground truth labels are available. Our goal is to derive sequential reject classifiers that reduces cost of measurement acquisition and error in the prediction(or testing) phase.

We show that this sequential reject classifier problem can be formulated as an instance of a Markov Decision Problem (MDP) when the class-specific probability models for the different sensor measurements are known. In this case the optimal sequential classifier can be cast as a solution to a Dynamic Program (DP). The DP solution is a sequence of stage-wise optimization problems, where each stage problem is a combination of the cost from the current stage and the cost-to-go that is carried on from later stages.

Nevertheless, class probability models are typically unknown; our scenarios produce high-dimensional sensor data (such as images). Consequently, unlike some of the conventional approaches Ji and Carin (2007), where probability models are first estimated to solve MDPs, we have to adopt a non-parametric discriminative learning approach. We formulate a novel multi-stage expected risk minimization (ERM) problem. This ERM formulation closely emulates limiting stage-wise-optimization suggested by the Dynamic Programming solution to the Markov Decision Problem(MDP). We solve this ERM problem at each stage by first factorizing the cost function into classification and rejection decisions. Then we transform reject decisions into a binary classification problem. Specifically, we show that the optimal reject classifier at each stage is a combination of two binary classifiers, one biased towards positive examples and the other biased towards negative examples. The disagreement region of the two then defines the reject region. We then approximate this empirical risk with a global surrogates. We present an iterative solution and demonstrate local convergence properties. The solution is obtained in a boosting framework. We then extend well-known margin-based generalization bounds Bartlett et al. (1998) to this multi-stage setting.
In our most recent work we develop further generalizations of the multi-stage decision framework. The system is no longer constrained to boosting classifiers. Any classification algorithm can be adapted to a sequential decision setting. We accomplish this by reducing the multi-stage ERM formulation to a series of supervised learning problems. Where at each stage, the reject and classification decisions can be learned by many standard supervised learning techniques (i.e. logistic regression, SVM, etc.). This formulation allows the system to handle multi-class decision tasks and utilize “black box” classifiers that are pre-programmed into a sensing modality. We derive bounds for generalization error for our sequential decision rules. In the binary classification setting our system turns out to be a Boolean fusion of binary decision functions. Using this insight, we derive an upper bound on the VC dimension of the multi-stage reject classifier. VC dimension is a measure of classifier complexity. We show that the VC dimension of a K-stage system grows as K log K times the maximum complexity of any stage; complexity of the system does not explode with the number of stages. This result is important because, for a fixed training phase performance, the generalization error is inversly proportional to the complexity of the system.

We tested our methods on synthetic, medical and explosives datasets. Our results demonstrate an advantage of multistage classifier thus resulting in cost reduction without significant sacrifice in accuracy.

Theory

We briefly present some of mathematics that govern much of our approach. Note that in our context the statistical models governing sensor measurements are unknown. Nevertheless, to gain insight, it is worth considering the scenario when such statistical models are indeed known. We consider a two-stage decision system for the sake of exposition. It turns out that in this case the Bayes optimal classification scheme is obtained by thresholding the likelihood function. In the formula on the next page, \( f(x) \), is the decision rule; \( y \) takes value one if the ground truth is that there is an explosive on the person; \( x \) is the features obtained at stage 1. In the explosives detection example this would be the measurements obtained by an X-ray scanner. The main difference between the conventional decision scheme and this problem is that the threshold here is data dependent and incorporates measurement costs. Specifically, the optimal decision at stage 1 is to reject (decide \( r \)) if the likelihood is sandwiched between two thresholds.

\[
f^1(x^1) = \begin{cases} 
+1, & P(y = 1|x^1) \geq 1 - \delta(x^1) \\
-1, & P(y = 1|x^1) \leq \delta(x^1) \\
r, & \delta(x^1) \leq P(y = 1|x^1) \leq 1 - \delta(x^1) 
\end{cases}
\]

The 2nd stage cost is given as follows:

\[
\tilde{\delta}(x^1) = \min_{f^2} E \left[ \text{error}(f^2, x^2, y) \mid x^1 \right] + \delta
\]

where \( \delta \) is the measurement cost for the 2nd stage. It turns out that we can split the optimal decision rule at stage 1 into two standard decision rules \( f_n \) and \( f_p \) as follows:

\[
f^1(x^1) = \begin{cases} 
f_p(x^1), & \text{if } f_n(x^1) = f_p(x^1) \\
r, & \text{if } f_n(x^1) \neq f_p(x^1) 
\end{cases}
\]

The classifier \( f_n \) is the decision rule for a standard Bayes optimal problem where the cost of missed detections is larger than the cost of false alarms. The bias in the costs depends on the measurement cost \( \delta \). For \( f_p \) the cost of false alarms are similarly chosen to be larger than the cost of missed detections. The optimal reject classifier is to reject examples to the second stage when the two classifiers do not agree.
Empirical Risk Minimization

In the empirical setting we do not have the likelihood models and have to learn decision rules based on training data: \((x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\). Our idea is to mimic the situation where models are known and find suitable surrogate data-dependent costs that account for the future stages. It turns out that this can be formulated in terms of empirical risk minimization problem. Specifically, for the two stage example, we use the following substitution:

\[
\tilde{\delta}_i = \delta + \mathbb{I}[f^2(x^2_i) \neq y_i]
\]

where the second term is an indicator function denoting whether or not the second stage is capable of classifying the example correctly. With this substitution the classifier for the first stage can be obtained by minimizing a suitable risk function:

\[
\frac{1}{N} \sum_{i=1}^{N} \left[ \mathbb{I}[f_p(x^1_i) \neq y_i] + \mathbb{I}[f_a(x^1_i) \neq y_i] \right] + \tilde{\delta}_i \mathbb{I}[f_p(x^1_i) \neq f_p(x^1_i)]
\]

The classifies \(f_a\) and \(f_p\) are the negatively and positively biased examples described before. To find the optimal classifier we can then use differentiable surrogates for indicator functions, and/or parametric classifiers in boosting framework.

Extension to Multi –Stages and Multi-Class setting

Similar to a two-stage setting, for each stage \(k\) we can define a risk functional. Here, the penalty for rejecting is the expected loss of the rest of the system (stages \(k+1\)... \(K\)), and the penalty for staying at the current stage is the performance of the current classifier \(f^k(\cdot)\). Again since the underlying probability models are not known, we form an empirical risk minimization for each stage:

\[
f^k(x^k) = \arg \min_{f \in \mathcal{F}^k} \frac{1}{N} \sum_{i=1}^{N} \tilde{R}_k(y_i, x^k_i, f, \tilde{\delta}^k_i)
\]

Next, we decompose decision \(f^k(\cdot)\) into (multi-class) classifier \(d^k(\cdot)\), its confidence \(\sigma(\cdot)\), and rejector \(g^k(\cdot)\).

\[
f^k(x^k) = \begin{cases} 
\delta^k(x^k), & f^k(x^k) = r \\
\sigma_{R}(x^k), & f^k(x^k) \neq y \land f^k(x^k) \neq r 
\end{cases}
\]

This decomposition allows us to alternatively solve for stage rejector while fixing the stage classifier and vice-versa. Each sub-problem for a stage rejector simplifies to a weighted supervised binary classification. To solve this sub-problem, one can utilize many existing techniques in the machine learning literature.

The entire system is trained by cyclical optimization. We pick a stage, hold the rest of the system constant and learn a decision at that stage using our classifier/rejector decomposition. Then we update cost-to-go for training examples for the rest of the system and move to learning the next stage. This procedure is repeated until the solution converges.
Results on Simulated Explosives Dataset

We now describe some of our results on a simulated explosives detection dataset. This dataset contains images taken of people wearing various simulated explosives devices. The imaging is done in three modalities: infrared (IR), passive millimeter wave (PMMW), and active millimeter wave (AMMW). All the images are registered. We extract many patches from the images and use them as our training data. A patch carries a binary label, it either contains a threat or is clean. IR and PMMW are the fastest modalities but also less informative. AMMW is slow since it requires raster scanning a person but it is the most useful. There are total of 1230 body images in the dataset. The basic inspection strategy is illustrated in Fig. 3.

![Figure 3: Illustration of Inspection Strategy consisting of three different scanning schemes.](image)

We also obtained ROC curves that highlight the advantages of using our inspection strategy over a centralized approach wherein the person is scanned with all the sensor modalities before any decision is made (Figure 4).

Note that at 47% reject rate we obtain the same detection/false alarm performance as a centralized system where all the sensor measurements are first obtained before any decision is made.

![Figure 4: ROC curves for different reject rates](image)

Results on Simulated Sequence of Sensors with Increasing Resolution

For this experiment, we consider a uniformly increasing cost structure. A sample using the 1st stage sensor incurs a cost of 1. To reach the second stage sensor the cost is 2 and so on. So for a 4-stage system, if a sample passes all four stages, it incurs a cost of 4. For an illustrative example, we convert a popular digit recognition

![Figure 5: Comparison of our method to myopic on the MNIST data. We construct four stages of increasing resolution by averaging the original digit images. The experiment demonstrates the advantage of our approach. Also note that the performance of a full resolution sensor can be achieved using a much lower resolution measurement.](image)
data, MNIST, into a four-stage decision system. We designate the full resolution 28x28 pixel image as the last stage. To simulate the first three stages of increasing sensor quality, we average the original image down to three resolution levels, 4x4, 7x7 and 14x14 pixels. The simulation demonstrates the advantage of our approach over a myopic strategy. The performance close to the centralized (best) strategy can be achieved with much lower average budget.

IV. FUTURE PLANS

Our research in these areas is breaking new ground. There are several interesting directions to pursue. The first direction is to complete the theory for the design and training of networks of classifiers where subsequent classifiers do not have guaranteed performance, where we model explicitly the processing time and try to optimize performance given a time budget. A second direction is extension of active learning principles to networks of classifiers. A third direction is evaluation of the theoretical results on sample data sets representative of homeland security applications. We attempted to work with the Sandia National Laboratory data set for whole-body scanning, and are currently in discussion with other industrial partners that may be able to provide representative data concerning scenarios of interest to DHS. As a final direction, we will investigate the potential use of these techniques to identify on-line samples for algorithm retraining in luggage inspection.

To extend our framework, we will consider fully adaptive sensor selection. In this setting, the order of the sensors is not fixed apriori, and the system will decide which sensor to use next based on the observed measurements.

V. LEVERAGING OF RESOURCES

One of the early students in this work, E. Rodriguez-Díaz, was supported in part by the National Science Foundation for his involvement in the research on sequential classifiers. These results have transitioned to incorporation in a medical device for colon and throat cancer detection using multispectral optical measurements.

VI. PROJECT DOCUMENTATION AND DELIVERABLES

A. Publications and Conference Presentations

1. Y. Chen, J. Qian, A New One-Class SVM for Anomaly Detection, ICASSP 2013
2. K. Trapeznikov, V. Saligrama, D. Castanon, Multi-Stage Classifier Design, Machine Learning, 2013
5. J. Qian, V. Saligrama, New Statistic in $p$-value Estimation for Anomaly Detection, IEEE SSP 2012

VII. REFERENCES


