LEARNING TO CLASSIFY WITH POSSIBLE SENSOR FAILURES

Tianpei Xie[†], *Nasser M. Nasrabadi*[⋆] *and Alfred O. Hero III*[†]

[†]Dept. of Electrical Eng., System, University of Michigan, Ann Arbor, MI 48109
[⋆] U.S. Army Research Lab., 2800 Powder Mill Road, Adelphi, MD, USA
[†] {tianpei, hero}@umich.edu, [⋆] nasser.m.nasrabadi.civ@mail.mil

ABSTRACT

In this paper, we propose an efficient algorithm to train a robust large-margin classifier, when corrupt measurements caused by sensor failure might be present in the training set. By incorporating a non-parametric prior based on the empirical distribution of the training data, we propose a Geometric-Entropy-Minimization regularized Maximum Entropy Discrimination (GEM-MED) method to perform classification and anomaly detection in a joint manner. We demonstrate that our proposed method can yield improved performance over previous robust classification methods in terms of both classification accuracy and anomaly detection rate using simulated data and real footstep data.

Index Terms— corrupt measurements, robust large-margin training, anomaly detection, maximum entropy discrimination

1. INTRODUCTION

Large margin classifiers, such as support vector machines (SVMs), have demonstrated good classification performance when the training data is representative of the test data [1, 2, 3]. However, in many real-world applications training data can suffer from corrupted measurements due to sensor failure. In such cases, unless one accounts for possible corruption of the training data, the performance of the classifier degrades significantly. This paper presents a new and effective approach to train classifiers with corrupted data.

There have been several approaches [4, 5, 6] to train classifiers in a manner that is robust to corrupted training data. Among these approaches, one common strategy is to use *Ramp Loss* [1, 7, 8], which explicitly limits the value of the maximal loss. The drawbacks of these Ramp-Loss-based approaches is that they do not provide a unified framework for joint anomaly detection and classification, and they are not capable of handling corrupted training samples.

The motivation behind our proposed algorithm is that corrupted data can be detected in the training set by using anomaly detection techniques [9] during the classifier training process. Such techniques are expressely designed to detect anomalies in order to attain the lowest possible false alarm and miss probabilities. In keeping with the non-parametric nature of the SVM classifier, we will focus on non-parametric anomaly detection schemes. Examples include minimal volume (MV) set anomaly detection [10, 11], and minimal entropy set anomaly detection [12, 13], etc. Among them, Hero et al. [12, 13] developed the Geometric Entropy Minimization (GEM) principle that estimates the MV set based on the k-nearest neighbor graph (k-NNG). The key contribution of this paper is incorporation of GEM anomaly detection into an SVM classifier under a non-parametric corrupt-data model.

Our proposed Geometric-Entropy-Minimization regularized Maximum Entropy Discrimination (GEM-MED) framework integrates large-margin training with anomaly detection using a Bayesian convex optimization framework. The Maximum Entropy Discrimination (MED) approach proposed by Jaakkola et al [14] performs Bayesian large margin classification using the maximum entropy principle. MED subsumes SVM as a special case. In this paper, we impose an *adaptive* large margin constraint for each of the sample instances. This constraint uses a Bayesian prior that is based on the GEM principle of anomaly detection. The resulting GEM-MED classifier effectively reduces the impact of anomalous samples on classification. We demonstrate superior performance on simulated data and on a real data set, containing human and human-leading-animal footsteps, collected in the field by acoustic sensors [15, 16, 17].

The outline of the paper is as follows: In Section **2**, the MED framework is introduced. In Section **3**, the GEM-MED framework is presented and a solution is proposed using variational inference. In Section **4**, experiment results based on synthesis data and real data are presented.

2. LARGE-MARGIN TRAINING WITH MED

Let the training data set be $\mathcal{D} \equiv \{(\mathbf{x}_n, y_n)\}_{n=1}^N$, where $\mathbf{x}_n \in \mathcal{R}^p$, $y_n \in \{-1, 1\}$. Assume the predictive distribution is log-linear, i.e. $\log p(y|\mathbf{x}, \Theta) \propto F(y, \mathbf{x}; \Theta) = \frac{1}{2} y(\mathbf{w}^T \mathbf{x} + b)$, where $\Theta = (\mathbf{w}, b)$ and $F(y, \mathbf{x}; \Theta)$ is the linear discriminative function parameterized by Θ. Denote the prior distribution of Θ as $p_0(\Theta)$. The *goal* for the *Maximum Entropy Discrimination* (MED) [14] is to learn a posterior distribution $p(\Theta|\mathcal{D})$, by solving an entropic regularized risk minimization problem

Acknowledgement: The research in this paper was partially supported by ARO grant WA11NF-11-1-103A1.

with prior $p_0(\Theta)$

$$\min_{p(\Theta|\mathcal{D})} \sum_n \left[1 - \mathbf{E}_{p(\Theta\,|\,\mathcal{D})}\{\Delta F(y_n, \mathbf{x}_n; \Theta)\}\right]_+ \quad (1)$$
$$+\mathbb{KL}\left(p(\Theta|\mathcal{D}) \| p_0(\Theta)\right),$$

where $[s]_+ = \max\{s, 0\}$. $\mathbb{KL}(p\|q)$ is the *Kullback-Leibler divergence* from distribution p to q, i.e. $\mathbb{KL}_{p(\Theta|\mathcal{D})}\left(p(\Theta|\mathcal{D})\|p_0(\Theta)\right) = \int_\Theta p(\Theta|\mathcal{D}) \log\left(\frac{p(\Theta|\mathcal{D})}{p_0(\Theta)}\right) d\Theta$ and $\Delta F(y_n, \mathbf{x}_n; \Theta) \equiv F(y_n, \mathbf{x}_n; \Theta) - F(y \neq y_n, \mathbf{x}_n; \Theta) = \log\left(\frac{p(y_n|\mathbf{x}_n, \Theta)}{p(y\neq y_n|\mathbf{x}_n, \Theta)}\right)$ is the log-odds that defines the classifier.

The first term in (1) is a hinge-loss that captures the large-margin principle underlying the MED prediction rule,

$$y^* = \text{argmax}_y \, \mathbb{E}_{p(\Theta|\mathcal{D})}\left[F(y, \mathbf{x}; \Theta)\right].$$

As in kernel SVM, we could extend the predictive distribution to a log non-linear function by kernelization [18, 19]. Specifically, a kernelized version of MED (1) is implemented by redefining the predictive distribution as $F(y, \mathbf{x}; \mathbf{w}) = \frac{1}{2} y \mathbf{w}^T \mathbf{\Phi}(\mathbf{x})$, where Φ is the prescribed feature map. Define the kernel function $K : \mathcal{R}^p \times \mathcal{R}^p \mapsto \mathcal{R}$ that satisfies $\langle \mathbf{\Phi}(\mathbf{x}_i), \mathbf{\Phi}(\mathbf{x}_j)\rangle = K(\mathbf{x}_i, \mathbf{x}_j)$, for $\forall \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}$. If we use a *Gaussian Process* [20] as the prior on $\mathbf{w}$, i.e. $p_0(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \mathbf{0}, \mathbf{I})$, kernel MED is obtained by solving (1) with $\Theta = \mathbf{w}$. The kernel MED approach is adopted in Sec. 4 but, for simplicity, we assume the log-linear model in Sec. 3.

Since the hinge loss in (1) is applied to *every* sample instance with equal weight, the training the MED classifier can be overly sensitive to anomalous samples. Therefore, it is desirable to extend MED to account for the possible presence of anomalous samples in the training set.

3. GEM REGULARIZED MED (GED-MED)

3.1. GEM regularized MED: model development

We develop a model that explicitly accounts for the possible presence of anomalous samples, by introducing a latent variable $\eta_n \in \{0, 1\}$ associated with each sample $\mathbf{x}_n$, where $\eta_n = 1$ means that $\mathbf{x}_n$ is nominal (uncorrupted), and $\eta_n = 0$ means that $\mathbf{x}_n$ is anomalous. Denote $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_N]^T$.

The *goal* is to learn the posterior joint distribution $p(\Theta, \boldsymbol{\eta}\,|\,\mathcal{D})$, given the prior $p_0(\Theta, \boldsymbol{\eta}) = p_0(\Theta)\, p_0(\boldsymbol{\eta}) = p_0(\Theta) \prod_n p_0(\eta_n)$. In analogy to the MED principle in (1), we propose a *regularized* MED framework to achieve this goal,

$$\min_{p(\Theta, \boldsymbol{\eta}\,|\,\mathcal{D})} \sum_n \mathbf{E}_{p(\Theta, \boldsymbol{\eta}\,|\,\mathcal{D})}\{\eta_n \left[1 - \Delta F(y_n, \mathbf{x}_n|\Theta)\right]_+\} \quad (2)$$
$$+C\, \mathcal{R}_{p(\Theta, \boldsymbol{\eta}\,|\,\mathcal{D})}(\boldsymbol{\eta}) + \mathbb{KL}(p(\Theta, \boldsymbol{\eta}\,|\,\mathcal{D}) \| p_0(\Theta)\, p_0(\boldsymbol{\eta})),$$

where $\mathcal{R}_{p(\Theta, \boldsymbol{\eta}\,|\,\mathcal{D})}(\boldsymbol{\eta})$ is the regularizer on $\boldsymbol{\eta}$ associated with $p(\Theta, \boldsymbol{\eta}\,|\,\mathcal{D})$, and $C > 0$ is the regularization parameter.

Note that the first term in (2) couples the quality of data, η_n, to the class prediction loss $\left[1 - \Delta F(y_n, \mathbf{x}_n|\Theta)\right]_+$. Compared with (1), this empirical risk is *adaptive* to the latent state η of the training samples, since a suspected anomalous instance (i.e. $\eta_n = 0$) will have no impact on the risk function.

The main contribution of this paper is the inclusion of the regularization term $\mathcal{R}_{p(\Theta, \boldsymbol{\eta}\,|\,\mathcal{D})}(\boldsymbol{\eta})$ into the MED framework. This term is estimated via GEM principle [12, 13] and it can be interpreted equivalently as the log of an empirical Bayesian prior on the quality indicator $\boldsymbol{\eta}$. In contrast to a data-independent prior $p_0(\boldsymbol{\eta})$, the proposed prior $\mathcal{R}_{p(\Theta, \boldsymbol{\eta}\,|\,\mathcal{D})}(\boldsymbol{\eta})$ depends on the empirical probability distribution of the training data $\mathcal{D}$. This is further discussed in Sec. 3.2.

The motivation for our approach is that the Ramp-Loss-based robust classification methods [1, 7, 8, 5, 6] fail to capture a crucial characteristic property of the anomalous data: such data tends to lie in the tail (low probability region) of the data set. As previous classification models do not explicitly account for the tails of the empirical distribution of the training data, one might expect significant classification performance improvement by incorporating such tail information via $\mathcal{R}_{p(\Theta, \boldsymbol{\eta}\,|\,\mathcal{D})}(\boldsymbol{\eta})$.

3.2. The construction of regularizer via GEM principle

To construct the regularization term $\mathcal{R}_{p(\Theta, \boldsymbol{\eta}\,|\,\mathcal{D})}(\boldsymbol{\eta})$ we follow the Geometric Entropy Minimization (GEM) [12, 13] principle. Specifically, GEM estimates Ω_β, where $\Omega_\beta = \arg\min_A\{H(A) : \int_A p(\mathbf{x})d\mathbf{x} \geq 1 - \beta\}$ is the minimal-entropy -set of false alarm level β, $H(A) = -\int_A \log p(\mathbf{x})\, p(\mathbf{x})d\mathbf{x}$ is the Shannon entropy of the density $p(\mathbf{x})$ over the region A. Given Ω_β, a sample $\mathbf{x}_n$ is declared anomalous if $\mathbf{x}_n \notin \Omega_\beta$. The decision rule is the Uniformly Most Powerful Test at level β when the anomalies are drawn from an unknown mixture of known nominal distribution $p(\mathbf{x})$ and uniform anomalous distribution[12]. As the training data distribution $p(\mathbf{x})$ is unknown, GEM approximates this decision rule by replacing Ω_β with an empirical estimate $\hat{\Omega}_\beta$ using the empirical distribution $\hat{p}(\mathbf{x}_n)$. Since η_n is the indicator function of the event $\mathbf{x}_n \notin \Omega_\beta$, GEM reduces to solving $\boldsymbol{\eta}^* = \text{argmin}_A\{ -\frac{1}{N}\sum_n \eta_n \log(\hat{p}(\mathbf{x}_n)) : \frac{1}{N}\sum_n \eta_n \geq 1 - \beta\}$. To approximate $\boldsymbol{\eta}^*$, as in [12, 13] a k-nearest neighbor graph approximation to Ω_β is constructed from the data set.

In our model GEM is implemented by using a bipartite K-point kNN graph (BP-kNNG) [13]. Specifically, we first split the training set into two parts using the class labels. The BP-kNNG anomaly detector of [13] is then implemented on each part $\mathcal{X}_c = \{x_n : y_n = c\}$, $c \in \{\pm 1\}$ independently. This results in partitioning the data set $\mathcal{X}_c$ into two parts $\{\mathcal{X}_{N_c}, \mathcal{X}_{M_c}\}$. For each sample, $\mathbf{x}_n \in \mathcal{X}_{N,y_n}$, its local entropy is estimated via $-\log(\hat{p}(\mathbf{x}_n)) = d\,\log(R_k(\mathbf{x}_n)) - \log\left(\frac{k-1}{M_c\, c_d}\right)$, where $R_k(\mathbf{x}_n)$ is the sum of k-nearest neighbor (kNN) distance from the target sample $\mathbf{x}_n$ to its M_c reference samples in $\mathcal{X}_{M_c}$; d is the intrinsic dimension of $\mathbf{x}_n$ and c_d is the volume of the unit ball in d dimensions. We estimate $-\log(\hat{p}(\mathbf{x}_m))$ for $\mathbf{x}_m \in \mathcal{X}_{M_c}$ in a similar manner. Then the set $\mathcal{E}_c \equiv \{-\log(\hat{p}(\mathbf{x}_n)) : y_n = c\}$ is arranged in ascending elements of the order, and we denote the sum of the first s_ρ smallest entropy values as S_c. The threshold s_ρ is set using the heuristic $s_\rho = \arg\max_k\{|-\log(\hat{p}(\mathbf{x}^{[k-1]})) + \log(\hat{p}(\mathbf{x}^{[k]}))|\}$, where $-\log(\hat{p}(\mathbf{x}^{[k]}))$ denotes the k-th smallest elements in $\mathcal{E}_c$.

Given S_c for each class $c \in \{-1, +1\}$, the *GEM regularization term* $\mathcal{R}_{p(\Theta,\boldsymbol{\eta}\,|\,\mathcal{D})}(\boldsymbol{\eta})$ is defined as

$$\mathcal{R}_{p(\Theta,\boldsymbol{\eta}\,|\,\mathcal{D})}(\boldsymbol{\eta}) \tag{3}$$
$$\equiv \sum_{c\in\{-1,+1\}} \left[\mathbf{E}_{p(\Theta,\boldsymbol{\eta}\,|\,\mathcal{D})} \left(- \sum_{n:y_n=c} \eta_n \frac{\log(\hat{p}(\mathbf{x}_n))}{N} \right) - S_c \right]_+ .$$

This term will impose a penalty for $\eta_n = 1$, when $\mathbf{x}_n \notin \hat{\Omega}_\beta$, since the value of S_c will make the summand of (3) positive.

Substituting (3) into (2), the *GEM regularized-MED* (GEM -MED) is obtained as

$$\min_{p(\Theta,\boldsymbol{\eta}\,|\,\mathcal{D})} \sum_n \mathbf{E}_{p(\Theta,\boldsymbol{\eta}\,|\,\mathcal{D})}\{\eta_n \left[1 - \Delta F(y_n, \mathbf{x}_n|\Theta)\right]_+\} \tag{4}$$
$$+C \sum_{c\in\{-1,+1\}} \left[\mathbf{E}_{p(\Theta,\boldsymbol{\eta}\,|\,\mathcal{D})} \left(- \sum_{n:y_n=c} \eta_n \frac{\log(\hat{p}(\mathbf{x}_n))}{N} \right) - S_c \right]_+$$
$$+\mathbb{KL}(p(\Theta,\boldsymbol{\eta}\,|\,\mathcal{D}) \| p_0(\Theta)\, p_0(\boldsymbol{\eta})),$$

where, as in (2), C is a regularization parameter.

3.3. Solving GEM-MED via variational inference

Similar to the form of $p(\Theta|\mathcal{D})$ in MED, $p(\Theta, \boldsymbol{\eta}|\mathcal{D})$ in GEM-MED (4) takes the form

$$p(\Theta, \boldsymbol{\eta}|\boldsymbol{\alpha}, \boldsymbol{\mu}, \mathcal{D}) = \frac{1}{Z(\boldsymbol{\alpha}, \boldsymbol{\mu})}\, p_0(\Theta) \exp(-\sum_n \eta_n \alpha_n \{1 - \Delta F_n(\Theta)\})$$
$$\times p_0(\boldsymbol{\eta}) \exp\left(- \sum_{c\in\{-1,+1\}} \mu_c \left[- \sum_{n:y_n=c} \eta_n \frac{\log(\hat{p}(\mathbf{x}_n))}{N} - S_c \right] \right)$$
$$\text{s.t.} \quad \mathbf{0} \preceq \boldsymbol{\alpha} \preceq \mathbf{1}, \quad \mathbf{0} \preceq \boldsymbol{\mu} \preceq C\mathbf{1},$$

where $\mathbf{1} = [1, \ldots, 1]^T$, $\Delta F_n(\Theta) \equiv \Delta F(y_n, \mathbf{x}_n|\Theta)$. $Z(\boldsymbol{\alpha}, \boldsymbol{\mu})$ is the *partition function* of the distribution and $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_N]^T$ and $\boldsymbol{\mu} = [\mu_{-1}, \mu_1]^T$ are dual variables associated with the hinge loss and regularizer in (4), respectively.

To identify the dual variables $\boldsymbol{\alpha}$ and $\boldsymbol{\mu}$, it is required to solve for the minimum of the log-partition function

$$\max_{\mathbf{0}\preceq\boldsymbol{\alpha}\preceq\mathbf{1}, \mathbf{0}\preceq\boldsymbol{\mu}\preceq C\mathbf{1}} -\log Z(\boldsymbol{\alpha}, \boldsymbol{\mu})$$
$$= -\log \sum_{\boldsymbol{\eta}\in\{0,1\}^N} \int_\Theta p(\Theta, \boldsymbol{\eta}|\boldsymbol{\alpha}, \boldsymbol{\mu},\ \mathcal{D}) d\Theta. \tag{5}$$

As (5) is concave in $\boldsymbol{\alpha}$ and $\boldsymbol{\mu}$ we propose a *projected stochastic gradient descent algorithm* (PSGD) [21] with Gibbs sampling. The procedure is summarized in **Algorithm** 1. As noted, we define $\pi_n \equiv p(\eta_{n,T} = 1|\Theta_T, \mathcal{D})$ as the anomaly score for each sample, where T is the final iteration of the PSGD algorithm.

4. EXPERIMENT

4.1. Simulated data

The first experiment is conducted on a simulated data set. For each class $c \in \{\pm 1\}$ samples are generated from the bivariate Gaussian distribution, with mean $\boldsymbol{m}_{-1} = (3, 3)$ and $\boldsymbol{m}_{+1} = -\boldsymbol{m}_{-1}$ and common covariance $\Sigma = \begin{bmatrix} 20 & 16 \\ 16 & 20 \end{bmatrix}$. Here $\Theta = (\mathbf{w}, b)$ has Gaussian prior in product form $p_0(\Theta) = \mathcal{N}(\mathbf{w};\, \mathbf{0}, \mathbf{I})\mathcal{N}(b; 0, \sigma_b^2)$. For each sample n, the prior on η_n is set to be equally likely: $p_0(\eta_n) = 1/2$.

The anomalies in the training set are drawn uniformly from a ring with an inner radius of R and outer radius $R+1$, where the value of R indicates the noise level in the corrupted training set [5]. We fix the size of the training set to be 100 for each class, with ratio of anomaly samples denoted as r_a. The test set contains 2000 samples from each class.

We first compare the classification accuracy of MED/SVM using *LibSVM* [22], Robust-Outlier-Detection (ROD) [5] with outlier parameter $\rho \in [0.01, 1]$ and GEM-MED, under noise level $R = [15, 35, 55, 75]$ and corruption rate $r_a = [0.2, 0.3, 0.4, 0.5]$. All the reported results are averaged over 50 runs. Fig 1(a) shows the mean and standard deviation of the test errors for these models versus various noise level R (with $r_a = 0.2$), and Fig 1(b) shows the test errors under different corruption rate settings (with $R = 55$). For ROD, only $\rho \in \{0.02, 0.2, 0.6\}$ are shown for simplicity, while $\rho = 0.02$ is the best for $\forall \rho \in [0.01, 1]$. In both experiments, as the noise level or the corruption rate increases, the training data become less representative of the test data and the difference between their distributions increases, which causes a significant increase of test error for MED/SVM method. While both ROD and GEM-MED limit the maximal loss values during training, and thus prevent over-fitting to anomalies, GEM-MED outperforms ROD as it incorporates the nonparametric

Algorithm 1 GEM regularized MED

Input: $\mathcal{D} \equiv \{(\mathbf{x}_n,\, \mathbf{y}_n)\}_{n=1}^N$, where $\mathbf{x}_n \in \mathcal{R}^p$, $\mathbf{y}_n \in \{\pm 1\}$. Prior distribution $p_0(\Theta)$, $p_0(\eta_n)$ and the upper bound $S_c, c \in \{\pm 1\}$.

1: **Initialize:** Set $\boldsymbol{\mu}_0 = \mathbf{0}$. $\boldsymbol{\alpha}_0$ is set by applying MED on $\mathcal{D}$

2: **for** $t = 1, \ldots, T$ or until converge **do**

3: Compute the gradient of log-partition function w.r.t $\boldsymbol{\alpha}_t$ and $\boldsymbol{\mu}_t$, respectively, as

$$\frac{\partial \log Z(\boldsymbol{\alpha}_t, \boldsymbol{\mu}_t)}{\partial \alpha_n} = \mathbb{E}_{p(\Theta,\boldsymbol{\eta}|\boldsymbol{\alpha}_t,\boldsymbol{\mu}_t,\mathcal{D})} \left[\eta_n \left\{\Delta F(y_n, \mathbf{x}_n; \Theta) - 1\right\}\right], \quad n = 1, \ldots, N,$$

$$\frac{\partial \log Z(\boldsymbol{\alpha}_t, \boldsymbol{\mu}_t)}{\partial \mu_c} = \mathbb{E}_{p(\Theta,\boldsymbol{\eta}|\boldsymbol{\alpha}_t,\boldsymbol{\mu}_t,\mathcal{D})} \left[\sum_{n:y_n=c} \eta_n \frac{\log(\hat{p}(\mathbf{x}_n))}{N} + S_c \right], \quad c \in \{-1, +1\},$$

where the expectation is approximated via Gibbs sampling with the conditional density $p(\boldsymbol{\eta}|\hat{\Theta}, \mathcal{D}) = \prod_{n=1}^N p(\eta_n|\hat{\Theta}, \mathcal{D})$ and $p(\Theta|\hat{\boldsymbol{\eta}}, \mathcal{D})$ computed explicitly.

4: Update α_n and μ_c via projected gradient descent, i.e.

$$\alpha_{n,(t+1)} = \text{proj}_{\{\alpha:\, 0\le\alpha\le 1\}} \left\{ \alpha_{n,t} - \varphi \frac{\partial \log Z(\boldsymbol{\alpha}_t, \boldsymbol{\mu}_t)}{\partial \alpha_n} \right\}, \quad n = 1, \ldots, N,$$

$$\mu_{c,(t+1)} = \text{proj}_{\{\mu:\, 0\le\mu\le C\}} \left\{ \mu_{c,t} - \psi \frac{\partial \log Z(\boldsymbol{\alpha}_t, \boldsymbol{\mu}_t)}{\partial \mu_c} \right\}, \quad c \in \{-1, +1\},$$

where $\text{proj}_{\{w:\, 0\le w\le C\}}\{w\} \equiv \min\left(\max(w, 0), C\right)$ defines the projection of w on the feasible set $\{w : 0 \le w \le C\}$ via clipping and ψ, $\varphi > 0$ define the learning rate.

5: **end for**

Output: Assign label for test sample $\mathbf{x}$ as

$$\hat{y} = \text{argmax}_y \int p(y|\mathbf{x},\ \Theta)\, p(\Theta|\mathcal{D}) d\Theta,$$

where $p(\Theta|\mathcal{D}) = \sum_{\boldsymbol{\eta}\in\{0,1\}^N} p(\Theta, \boldsymbol{\eta}\,|\,\mathcal{D})$ is computed via marginalization. Also obtain the posterior on $\boldsymbol{\eta}$ at the final iteration of step 4, $\{\pi_n \equiv p(\eta_{n,T} = 1|\Theta_T, \mathcal{D})\}$, as the anomaly scores.

prior $\mathcal{R}(\boldsymbol{\eta})$ that is adaptive to anomalies in the training set, as opposed to ROD, which relies on the predefined tuning parameter ρ.

We then evaluate the efficiency of anomaly detection for various RODs and GEM-MED, under the fixed corruption rate 0.2. The $\pi_n's$ in GEM-MED and RODs are used as anomaly scores and are placed in ascending order. We compute the precision and recall using this ordering, averaging over 50 runs. Fig.2(a) plots the precision versus recall curve for various RODs and GEM-MED, with a snapshot of a typical result illustrated in Fig.2(b). As seen in both figures, the anomaly score given by GEM-MED provides more relevant information about the true anomalies, compared to that given by RODs. This is due to the additional GEM-based regularizer $\mathcal{R}(\boldsymbol{\eta})$ in GEM-MED, which captures the characteristic of an anomalous sample based on the relative entropy of the region in which it resides. GEM-MED, thus, has better performance in terms of the efficiency and accuracy of anomaly detection than does ROD.

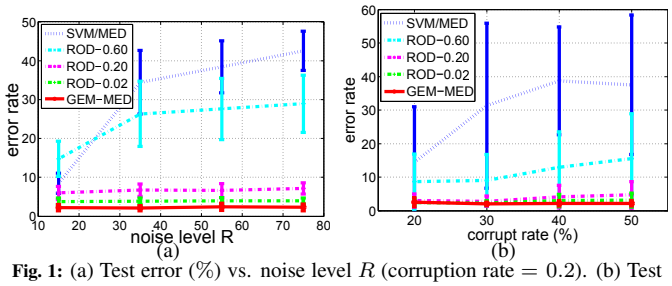

Fig. 1: (a) Test error (%) vs. noise level R (corruption rate $= 0.2$). (b) Test error (%) vs. corrupt rate ($R = 55$) on simulated data. GEM-MED outperforms both MED/SVM and ROD for various ρ in classification accuracy, when either noise level or corrupt rate increases .

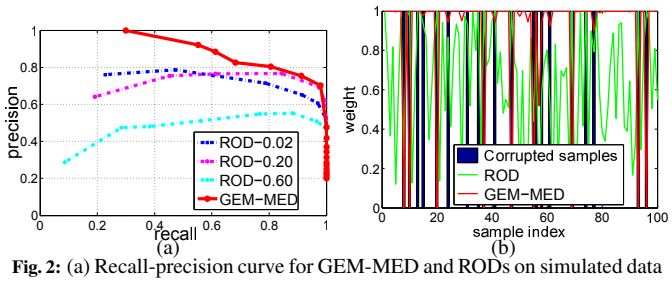

Fig. 2: (a) Recall-precision curve for GEM-MED and RODs on simulated data (corruption rate $= 0.2$). (b) Illustration of anomaly score π_n for GEM-MED and ROD-0.2. GEM-MED surpluses ROD in anomaly detection.

4.2. Footstep classification data set

We perform experiments on ARL-Footstep multisensor data set [17, 16, 23], where the task is to discriminate between human footsteps and human-leading animal footsteps. The footstep data was collected via four well-synchronized acoustic sensors (labeled as Sensor 1,2,3,4) in a natural environment, where the environmental noise and multiple sensor failures corrupted the acoustic recordings. It involves 84 human subjects and 66 human-animal subjects. We randomly select 25 subjects from each class as the training set, with the rest designated as the test set. In the preprocessing step, footsteps are detected, extracted and segmented before a 200-dimensional mel-frequency cepstral coefficients (MFCCs) vector is computed for each segment. We then apply PCA to reduce the dimensionality from 200 to 50, as in [17, 23]. For multiple D sensors the augmented feature of dimension $50D$ is constructed via feature concatenation.

In these experiments, we apply kernel MED with the Gaussian kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|_2^2)$; the kernel parameter $\gamma > 0$ is tuned via 5-fold-cross-validation. A Gaussian Process $\mathcal{N}(\mathbf{w}; \mathbf{0}, \mathbf{I})$ is used as prior on $\mathbf{w}$. For each sample n, the prior on η_n is set to $p_0(\eta_n) = 1/2$.

Table 1 shows the classification accuracy for four individual sensors and the combination of all four sensors with kernel MED, ROD for $\rho \in [0.01, 1]$ and GEM-MED and Table 2 shows their respective anomaly detection accuracies. For ROD only $\rho = 0.02$ and $\rho = 0.20$ are shown, while $\rho = 0.20$ is the best for $\forall\rho \in [0.01, 1]$. It is seen that the GEM-MED method outperforms all of the ROD-ρ algorithms and also outperforms kernel MED in classification accuracy for sensor 1,2,4 and it demonstrates significant improvement in detection for sensor 1,3,4. Note that ROD-0.2 has higher detection accuracy than GEM-MED in sensor 2, since many anomalous samples in this sensor reside in the high density region of the data set, which violates the sparse anomaly assumption underlying GEM. For combined sensors, as most of the anomalies based on the joint feature representation reside in the high entropy region of the data set, GEM-MED is able to successfully detect most of the anomalous samples.

Classification Accuracy (%) mean $\pm$ standard error				
sensor no.	kernel MED	ROD-0.02	ROD-0.2	GEM-MED
1	71.1 ± 5.3	73.7 ± 3.7	76.0 ± 2.5	$\mathbf{78.4 \pm 3.3}$
2	62.3 ± 10.2	71.5 ± 7.3	76.5 ± 5.3	$\mathbf{82.1 \pm 3.1}$
3	60.0 ± 13.1	63.2 ± 5.4	$\mathbf{67.6 \pm 4.2}$	66.8 ± 4.5
4	58.4 ± 8.2	71.8 ± 7.2	73.2 ± 4.2	$\mathbf{80.1 \pm 3.1}$
1,2,3,4	78.6 ± 5.1	79.2 ± 3.7	79.8 ± 2.5	$\mathbf{84.0 \pm 2.3}$

Table 1: Classification accuracy with different sensors, with the best performance shown in **bold**.

Anomalous Detection Accuracy (%) mean $\pm$ standard error			
sensor no.	ROD-0.02	ROD-0.2	GEM-MED
1	30.2 ± 1.3	59.0 ± 3.5	$\mathbf{70.5 \pm 1.3}$
2	23.5 ± 2.6	$\mathbf{63.5 \pm 2.8}$	63.4 ± 2.5
3	5.3 ± 1.4	48.1 ± 3.3	$\mathbf{72.8 \pm 1.5}$
4	22.8 ± 3.2	65.2 ± 4.2	$\mathbf{88.1 \pm 2.1}$

Table 2: Anomalous detection accuracy with different sensors, with the best performance shown in **bold**.

5. CONCLUSION

In this paper we propose the GEM-MED algorithm that provides a unified optimization framework for classification and anomaly detection. We demonstrate its performance advantages in terms of both classification accuracy and detection rate on a simulated data set and a real footstep data set, as compared to the anomaly-blind Ramp-Loss-based classification method.

6. REFERENCES

[1] Peter L Bartlett and Shahar Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *The Journal of Machine Learning Research*, vol. 3, pp. 463–482, 2003.

[2] Olivier Bousquet and André Elisseeff, "Stability and generalization," *The Journal of Machine Learning Research*, vol. 2, pp. 499–526, 2002.

[3] Bernhard Schölkopf and Alexander J Smola, *Learning with kernels*, The MIT Press, 2002.

[4] Qing Song, Wenjie Hu, and Wenfang Xie, "Robust support vector machine with bullet hole image classification," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 32, no. 4, pp. 440–448, 2002.

[5] Linli Xu, Koby Crammer, and Dale Schuurmans, "Robust support vector machine training via convex outlier ablation," *AAAI*, vol. 6, pp. 536–542, 2006.

[6] Lei Wang, Huading Jia, and Jie Li, "Training robust support vector machine with smooth ramp loss in the primal space," *Neurocomputing*, vol. 71, no. 13, pp. 3020–3025, 2008.

[7] Nir Krause and Yoram Singer, "Leveraging the margin more carefully," *Proceedings of the twenty-first international conference on Machine learning*, p. 63, 2004.

[8] Llew Mason, Jonathan Baxter, Peter L Bartlett, and Marcus Frean, "Functional gradient techniques for combining hypotheses," *Advances In Neural Information Processing Systems*, pp. 221–246, 1999.

[9] Varun Chandola, Arindam Banerjee, and Vipin Kumar, "Anomaly detection: A survey," *ACM Computing Surveys (CSUR)*, vol. 41, no. 3, pp. 15, 2009.

[10] Bernhard Schölkopf, Robert C Williamson, Alex J Smola, John Shawe-Taylor, and John C Platt, "Support vector method for novelty detection.," *Advances In Neural Information Processing Systems*, vol. 12, pp. 582–588, 1999.

[11] Clayton D Scott and Robert D Nowak, "Learning minimum volume sets," *The Journal of Machine Learning Research*, vol. 7, pp. 665–704, 2006.

[12] Alfred O Hero, "Geometric entropy minimization (GEM) for anomaly detection and localization," *Advances in Neural Information Processing Systems*, pp. 585–592, 2006.

[13] Kumar Sricharan and Alfred Hero, "Efficient anomaly detection using bipartite k-NN graphs," *Advances in Neural Information Processing Systems*, pp. 478–486, 2011.

[14] Tommi Jaakkola, Marina Meila, and Tony Jebara, "Maximum entropy discrimination," *Advances in Neural Information Processing Systems*, 1999.

[15] Thyagaraju Damarla, Asif Mehmood, and James Sabatier, "Detection of people and animals using non-imaging sensors," *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*, pp. 1–8, 2011.

[16] Thyagaraju Damarla, "Seismic and ultrasonic data analysis for characterizing people and animals," *SPIE Defense, Security, and Sensing*, 2012.

[17] Po-Sen Huang, Thyagaraju Damarla, and Mark Hasegawa-Johnson, "Multi-sensory features for personnel detection at border crossings," *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*, pp. 1–8, 2011.

[18] Jun Zhu, Ning Chen, and Eric P Xing, "Infinite latent SVM for classification and multi-task learning," *Advances in Neural Information Processing Systems*, pp. 1620–1628, 2011.

[19] Tony Jebara, "Multitask sparsity via maximum entropy discrimination," *The Journal of Machine Learning Research*, vol. 12, pp. 75–110, 2011.

[20] CE Rasmussen and CKI Williams, "Gaussian processes for machine learning," *Adaptive computation and machine learning*, 2006.

[21] Dimitri P Bertsekas, *Nonlinear programming*, Athena Scientific, 1999.

[22] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 27, 2011.

[23] Nam H Nguyen, Nasser M Nasrabadi, and Trac D Tran, "Robust multi-sensor classification via joint sparse representation," *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*, pp. 1–8, 2011.