objective is to find outliers that form a structure in data and that negatively impact decision boundary placement when training a model. We illustrate the simplest example using two data sets and later generalize to arbitrarily large number of sets. At each iteration of the procedure, we remove the 'most spurious' sample from the entire training set. To quantify spuriousness, we introduce a divergence based cost:

$$D_{global} = \sum_{i=1}^{N}\sum_{j=1}^{N} D(X_i||X_j), \tag{3}$$

where D is some divergence estimator and D_{global} is global divergence. Thus, our goal can be restated as minimizing global divergence, We chose the Renyi estimator for purposes of consistency, unless otherwise noted.

$$\text{Renyi } D_\alpha(P||Q) = \frac{1}{\alpha - 1} log \sum_{i=1}^{n} p_i^\alpha q_i^{1-\alpha} \tag{4}$$

D_α is strictly non-negative for $\alpha > 0$ and minimized when $P = Q$. Spurious samples misalign P and Q, thus samples with a large contribution to D_α are more likely to be spurious. If only the common structure remains, then we will not be able to improve $D_\alpha(P||Q)$.

Experimental Results

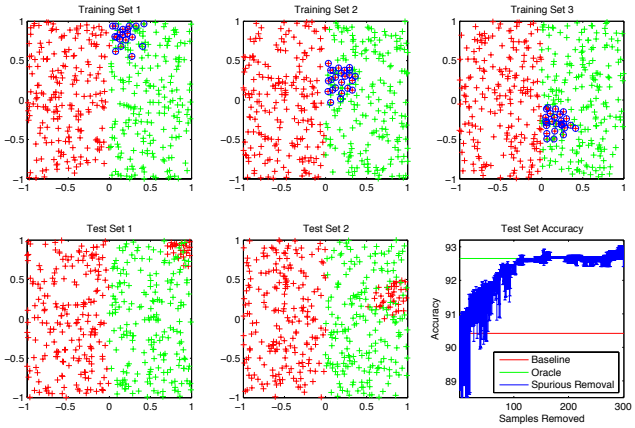

Figure 1: Spurious samples (circled in blue) removed in the training sets (top row) to retrieve the common structure. Test set accuracy during spurious sample removal (bottom right).

The artificial data sets in Fig. 1 illustrate how spurious samples negatively affect the placement of a linear SVM decision boundary for a binary classification task. We consider an oracle model trained on samples from the common distribution only (no spurious points). On the other hand, there is the baseline model, which is the result of training a linear SVM on all the data including all spurious samples. The presence of spurious samples shifts this linear decision boundary slightly, thus, the baseline divides the classes in a way which misclassifies some samples from the default distribution, decreasing the accuracy compared to the oracle.

We trained a model after each iteration of the greedy spurious sample removal to illustrate its effect. Then, we bootstrapped to entire process to obtain an average accuracy and a 95% confidence interval, shown in Fig 1. We found that, as we removed more samples, the clipped model performance approached the oracle, with tighter confidence intervals, thus the removal of spurious samples is indeed beneficial.

Now, let us consider a nuclear threat detection system, built for determining whether a vehicle that passes through customs emits signatures consistent with radioactive material. In Figure 2, we depict the most informative 2D projection, where a non-trivial density mismatch manifests for datasets generated with different simulation parameters. Threats are shown in red, normal samples shown in green.

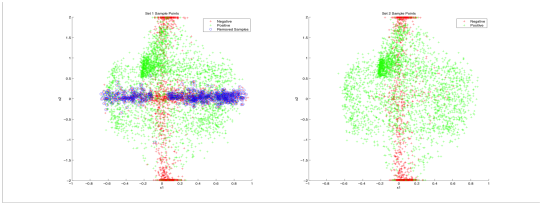

Figure 2: Nuclear threat datasets DS1 (left) and DS2 (right).

Figure 2 shows the blue circled spurious samples removed. The baseline we used (M_0) is trained on all data. Our approach produces a clipped version of DS1 which we added to DS2 to obtain the alternative model M_1. We test M_0 and M_1 on all other datasets. Additionally, we enhance our approach with the use of a gating function. That is, the model to be used in classification is determined by picking the model (M_0 or M_1) with the smallest Renyi divergence to the test set. We refer to this gated model as M_2. The justification for this is that some testing datasets can have spurious samples that are close enough to the ones in the original datasets, so it makes sense to use these samples, when beneficial. The gated version outperforms the other two as it benefits from sample removal when the incoming datasets do not have spurious samples, as shown in Table 1.

Table 1: Comparison of accuracy for a model using all the data (M_0), a clipped model (M_1) and the gated model (M_2)

	Sets resembling DS1	Sets resembling DS2
Acc M0	57.3692	89.4015
Acc M1	57.3197	89.4125
Acc M2	**57.3692**	**89.4125**

References

Breunig, M. M.; Kriegel, H.-P.; Ng, R. T.; and Sander, J. 2000. LOF: identifying density-based local outliers. In *ACM Sigmod Record*, volume 29, 93–104. ACM.

Lee, J.; Gilad-Bachrach, R.; and Caruana, R. 2013. Using multiple samples to learn mixture models. In *Advances in Neural Information Processing Systems*, 324–332.

Onderwater, M. 2010. Detecting unusual user profiles with outlier detection techniques.

Zou, J. Y.; Hsu, D.; Parkes, D. C.; and Adams, R. P. 2013. Contrastive Learning Using Spectral Methods. In *Advances in Neural Information Processing Systems*, 2238–2246.