Leveraging Common Structure to Improve Prediction across Related Datasets

Matt Barnes
mbarnes1@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA 15213

Nick Gisolfi
Carnegie Mellon University
Pittsburgh, PA 15213
ngisolfi@cmu.edu

Madalina Fiterau
mfiterau@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA 15213

Artur Dubrawski
awd@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

In many applications, training data is provided in the form of related datasets obtained from several sources, which typically affects the sample distribution. The learned classification models, which are expected to perform well on similar data coming from new sources, often suffer due to bias introduced by what we call 'spurious' samples – those due to source characteristics and not representative of any other part of the data. As standard outlier detection and robust classification usually fall short of determining groups of spurious samples, we propose a procedure which identifies the common structure across datasets by minimizing a multi-dataset divergence metric, increasing accuracy for new datasets.

Problem statement

Often, the data available for learning is collected from different sources, making it likely that the differences between these groups break typical assumptions such as the samples being independent and identically distributed. It is often the case that data from each source exhibits certain particularities - for instance, medical intervention might differ between hospitals, monitoring equipment might introduce systematic noise and patient vital signs will definitely vary by individual. We will refer to samples which are specific to the collection source as *'spurious'*, as they are not representative of data for which the model will be used.

The focus of this paper is the identification and removal of spurious samples from related training sets collected from multiple sources, with the objective of improving model performance on data from sources yet unobserved, under the assumption that the majority of samples come from a distribution common across all sources. This problem is an issue of importance in a multitude of scenarios, including: recommender systems, clinical models which need to be applied to new patients; prediction on census data when the collection was limited to a subset of locations.

In practice, we observe that a superior training set can be constructed using the most representative samples across datasets, by simply withholding groups of dataset-specific samples that are substantially different from the common distribution. We introduce a *'clipping'* procedure which removes samples from a dataset such that a model learned on it is more representative of the other available datasets. The 'clipped model' we obtain has improved accuracy compared to the standard model trained on all the data.

Assume the training data is given in the form of N datasets, each coming from a different source: $X_i \in \mathbb{R}^{n_i \times m}$, $Y_i \in \{0,1\}^{n_i}$, where $i \in \{1 \dots N\}$. We will refer to the r^{th} sample of dataset i as $(x_{i,r}, y_{i,r})$. Each sample is drawn from a distribution p_i as follows:

$$p_i(x,y) = \begin{cases} p^0(x,y) & \text{with prob. } q_i^0(y) \\ p_i^1(x,y) & \text{with prob. } 1 - q_i^0(y) \end{cases} \tag{1}$$

The distribution p^0 is common to all datasets, whereas the distributions p_i^1 for $i \in \{1 \dots N\}$ are responsible for the spurious samples. If the sample distribution was identical across datasets, then $q_i^0(y) = 1$, that is, there would be no spurious samples. By removing the spurious samples from the distributions $p_i^1(x,y)$, we have the opportunity to improve classification accuracy on samples drawn from p^0. We use the notation $(\bar{X}_i, \bar{Y}_i)$ to identify only the samples of dataset i which belong to the common distribution p^0. Also, let $h_{(X,Y)}$ be the classifier from the hypothesis class $\mathcal{H}$ which is learned from the training samples (X,Y). We will use the symbol $\frown$ to denote dataset stacking (concatenation). Within the previously-established framework, we have that, for a test set $(X_t, Y_t) \sim p^0$:

$$\begin{aligned} &mean(I[h_{(\bar{X}_1,\bar{Y}_1)\frown \dots \frown(\bar{X}_N,\bar{Y}_N)}(X_t) \neq Y_t]) \leq \\ &mean(I[h_{(X_1,Y_1)\frown \dots \frown(X_N,Y_N)}(X_t) \neq Y_t]) \end{aligned} \tag{2}$$

The key notion of considering the source of each outlier sets our works apart from standard outlier removal techniques such as Robust Mahalanobis Distance-based detection, Local Outlier Factor (Breunig et al. 2000) and Local Reconstructive Weights (Onderwater 2010). Some research does focus on leveraging different training sets, for instance Zou et. al (Zou et al. 2013) proposed using multiple data sets to improve performance of HMMs in contrastive learning. Lee et.al (Lee, Gilad-Bachrach, and Caruana 2013) recovered underlying structure, assuming the presence of several samples generated from the same underlying distributions with different mixing weights. However, our method is designed to work under more general settings.

Spurious sample removal procedure

Our approach builds on the intuitive use of density estimation in outlier detection, while using the information provided by simultaneously analyzing multiple data sets. The 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 function

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

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

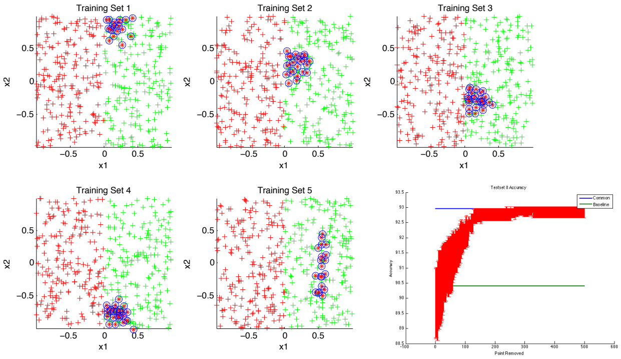

Figure 1: The samples circled in blue (spurious) are removed to retrieve the common structure of the related datasets. Accuracy during spurious sample removal (bottom right).

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 rotates 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 the purpose of determining whether a vehicle that passes through customs emits signatures consistent with radioactive material. Because of the high cost in collecting training data – most notably for the threat class – physicists generate synthetic data for training. This is an iterative process, with significant density mismatch between subsequent sets of data. In Figure 2, we depict the most informative 2D projection, for datasets sets generated at different stages. Threats are shown in red, normal samples shown in green.

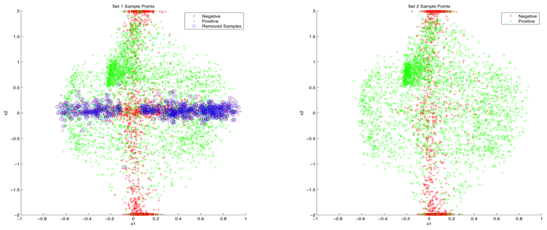

Figure 2: Nuclear threat datasets DS1 (left) and DS2 (right).

Figure 2 shows the blue circled spurious samples successfully removed. The baseline which 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.