Spatial Active Learning For Cost-Effective Sensing and Feature Extraction

Matthew Berger* MATTHEW.BERGER.1@US.AF.MIL
Lucas Magee[†] MAGEE.L@HUSKY.NEU.EDU
Lee M. Seversky*, Eric Heim* {LEE.SEVERSKY, ERIC.HEIM.1}@US.AF.MIL
*Air Force Research Laboratory, Information Directorate, Rome, NY USA
[†]Northeastern University, Department of Computer Science, Boston, MA USA

1. Introduction

Pool-based active learning methods have historically been used as a means to reduce the amount of supervised data necessary to learn accurate inference models. However, such methods assume that the pool of unsupervised instances from which the active learner chooses is known a priori. More specifically, the active learner must have access to the features of these instances that will be used for learning. In many practical learning scenarios there is a cost in obtaining these features. For instance, feature extraction methods may be computationally expensive, or simply observing instances can be expensive when state-of-the-art sensors are required. Hence pool-based active learning methods, while cost effective in supervision, can be deemed data inefficient due to the need for a pool of fully descriptive data instances.

Fortunately, data instances often have alternative, low-cost side information that can be used to reason about their relationships. In this work, we focus on multi-class classification tasks where instances are *spatial* in nature – each data instance is associated with a physical location on the Earth. We assume a scenario where data instances are acquired via spatial queries – this is common in visual recognition tasks where data sources such as Flickr, Google Maps, and Google Streetview serve images via geographic coordinates. Our main observation is that the uncertainty in the labels of unsupervised data instances is smooth with respect to their physical locations. Thus uncertainty in an instance's label can be approximated using the spatial relationships it has with labeled instances. This allows us to informatively select data instances to collect and compute features, leading to a low-cost, data-efficient active learning scheme. In the remainder of this paper, we formalize our method for spatial active learning, and show its strengths for data-scarce classification of satellite imagery and aerial LiDAR data.

2. Methodology

We begin by establishing some notation. Let $\mathcal{S} = \{s_1, ...s_n\}$ be a pool of two-dimensional coordinates corresponding to unlabeled instances. Let $\mathcal{X} = \{x_1, ...x_n\}$ be their features used to learn a classifier, and $\mathcal{D} = \{(x_1, y_1), ..., (x_n, y_n)\}$ be the full set of labeled, fully-featured data, where $y_1, ..., y_n \in Y = \{1, ..., C\}$ is the set of labels. In this work we use multinomial logistic regression as our model, which defines the conditional probability of a class c given an instance x_i as:

$$p(Y = c|x_i) = \frac{1}{1 + e^{-w_c^{\mathrm{T}} x_i}}, \tag{1}$$

where w_c are the logistic regression parameters for class c, and $W = \{w_1, ..., w_C\}$ are the parameters for the full model. With this definition, entropy can be used as a means to measure the uncertainty in the label of an instance:

$$H(Y|x_i) = -\sum_{c=1}^{C} p(Y = c|x_i) \log(p(Y = c|x_i)) \tag{2}$$

Using these entropy scores, the instance for which the logistic model is most uncertain about can be chosen:

$$x^* = \underset{x_i \in \mathcal{X}}{\operatorname{argmax}}\, H(Y|x_i) \tag{3}$$

A common technique in active learning, called uncertainty sampling, iteratively selects instances to be labeled at a given round t by finding x^* given the current logistic model W_{t-1}, posing it to a source of supervision for its true label, and retraining the model with this labeled instance.

In our learning setting, the active learner does not have access to the full pool of fully-featured instances $\mathcal{X}$. Instead it has access to $\mathcal{S}$, and can select instances for which features are constructed. We call our algorithm for this setting Spatial Active Learning (SAL). We assume that there exists a map $f : \mathcal{S} \mapsto \mathcal{X}$ that maps spatial coordinates to features, and that f is *smooth*: points whose physical coordinates are close implies that their features are a small distance apart – see Figure 1 (left) for an illustration. This manifests as smoothness of H, namely $H(Y|f(s))$ varies smoothly as a function of $s \in S$, highlighted in Figure 1 (right). This allows us to estimate entropy at any point $s \in S$ from a sparse set of known entropy values.

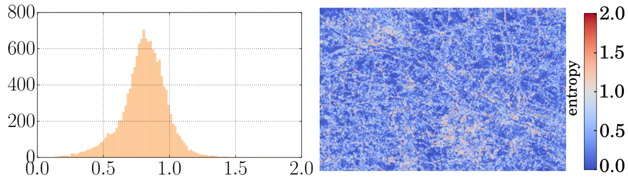

Figure 1: Left: For each instance in the CNY dataset, the ratio of the average distance of its features to its four nearest spatial neighbors, and the average distance to randomly sampled instances, as a histogram over all data (<1 indicates spatial neighbors are close in full-featured domain). Right: Example approximate entropy field.

Our active learning scenario at each round t is characterized by four components: $\mathcal{S}$, W_{t-1}, $\mathcal{X}_{t-1} \subset \mathcal{X}$, and $\mathcal{D}_{t-1} \subset \mathcal{D}$, the last two being a pool of instances for which features were computed and labeled data, respectively. Initially, SAL constructs an initial pool $\mathcal{X}_0$ by randomly sampling from $\mathcal{S}$ and applying f. All of these instances receive labels to construct $\mathcal{D}_0$. From there, SAL performs subsequent active selection rounds consisting of two stages. The first stage is exactly the uncertainty sampling technique described above on the reduced pool of instances $\mathcal{X}_{t-1}$. In the second stage, SAL adds to $\mathcal{X}_{t-1}$ by first constructing an approximate entropy field over $\mathcal{S}$. It does this by directly computing the entropy of the instances in $\mathcal{X}_{t-1}$ with respect to W_{t-1}, and interpolating them in the spatial domain. We employ radial basis function (RBF) interpolation over the 2D domain, where $H(Y|x_i)$ for each feature $x_i \in \mathcal{X}_{t-1}$ defines a scalar constraint at its corresponding 2D position $s_i \in S$. We use the polyharmonic spline radial basis function $\phi(a) = a^2 \log(a)$ as the RBF (Buhmann, 2003), leading to an approximation of H defined over S at round t, denoted $\hat{H}_t(Y|s)$ for $s \in \mathcal{S}$.

With this approximate measure of uncertainty, SAL actively selects instances in the spatial domain to form $\mathcal{X}_t$:

$$\mathcal{X}_t = \mathcal{X}_{t-1} \cup \left\{ f\left(\underset{s_i \in \mathcal{S}}{\operatorname{argmax}}\, \hat{H}_t(Y|s_i) \right), f(\operatorname{rand}(\mathcal{S})) \right\} \quad (4)$$

The first instance is the feature constructed from the spatial coordinate with highest approximate entropy. The second is the feature constructed from a randomly selected coordinate. These two elements are a trade-off between exploiting the instance that is estimated to be most informative, and randomly exploring the spatial domain. In summary, after an initialization round, SAL performs rounds of standard uncertainty sampling according to (3) from $\mathcal{X}_{t-1}$, followed by constructing $\mathcal{X}_t$ according to (4).

3. Experiments

We apply our method on two data sets of different modalities to highlight the strengths of SAL. The first data set is aerial 3D LiDAR data scanned over California. The data set is formed by tiling a given spatial region, where each tile has a physical location and a feature derived via a bag-of-words model defined on estimated surface normal variation. Each tile has a binary label, taking on whether or not the tile contains man-made structures. The second data set is satellite imagery taken over a region in central New York that is similarly tiled. We use CNN features from the AlexNet model trained on ImageNet (Krizhevsky et al., 2012), using activations from the penultimate layer of the network. Each tile contains a label consisting of farmlands, water, commercial building, residential building, or industrial building. As a baseline, we compare to methods that similarly perform uncertainty sampling for active learning, but add features to the pool *randomly* at every round, as studied in (Ertekin et al., 2007). We vary the amount of samples randomly added to the pool – note that SAL only adds 2 samples to the pool at every round. For each data set we evaluate the methods by comparing the classification error over each round, taken as the mean error over 20 trials, with each trial randomly bootstrapped with a different initial pool.

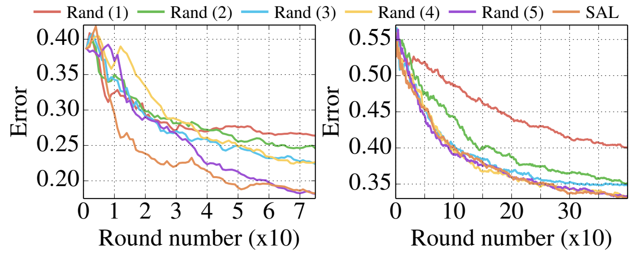

Figure 2: Classification error (left: LiDAR, right: imagery)

See Figure 2 for the results. We observe that SAL consistently outperforms the baseline when 2 samples are randomly added, and even outperforms the baseline for larger budgets. Since our method adds samples to the pool that are likely to have high entropy, our method has a better chance of selecting high uncertainty samples when compared to random sampling. This highlights the fact that entropy can be reliably estimated in such data-starved environments, where by using the underlying spatial domain we can informatively select data points to add to the pool for active learning.

References

Buhmann, Martin D. *Radial Basis Functions: Theory and Implementations*, volume 12. Cambridge University Press, 2003.

Ertekin, S., Huang, J., Bottou, L., and Giles, L. Learning on the border: active learning in imbalanced data classification. In *CIKM*, pp. 127–136, 2007.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. Imagenet classification with deep convolutional neural networks. In *NIPS*, pp. 1097–1105, 2012.