Active Learning for Informative Projection Retrieval

Madalina Fiterau
mfiterau@cs.cmu.edu
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh PA 15213

Artur Dubrawski
awd@cs.cmu.edu
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh PA 15213

Abstract

We introduce an active learning framework designed to train classification models which use informative projections. Our approach works with the obtained low-dimensional models in finding unlabeled data for annotation by experts. The advantage of our approach is that the labeling effort is expended mainly on samples which benefit models from the considered hypothesis class. This results in an improved learning rate over standard selection criteria for data from the clinical domain.

Introduction

Many applications of decision support systems require that their users comprehend the labeling suggestions. We have recently developed a method, called *Regression for Informative Projection Retrieval* (RIPR) (Fiterau and Dubrawski 2012), which learns compact models consisting of informative low-dimensional axis-aligned projections. The RIPR models reveal predictive structure in data, provided that it exists. We introduce an approach which recovers informative projections in the more challenging active learning setting. Our innovation is the *ActiveRIPR* framework, which enables active selection of yet unlabeled data which specifically targets the construction of accurate RIPR models.

Our framework selects samples to be labeled based on the relevant dimensions of the current classification model, trained on previously annotated data. The labeling effort is thus shifted to labeling samples that specifically benefit the low-dimensional models of our chosen hypothesis class. We enhance standard active selection criteria using the information encapsulated by the trained model. Thus, high accuracy is achieved faster than with standard sampling techniques, reducing the annotation effort exerted by domain experts. An added benefit is that the informative projections highlight structure that experts should be aware of and are available during labeling in addition to the full-featured data.

Despite the popularity of active learning, there is little research in the direction of using it for feature selection. Relevant contributions include (Bilgic 2012), introducing a methods which performs PCA on the data during the query selection process, selecting the features with the largest eigenvalues and performing L_2 regularization on them. In their setup, unlike here, the allowed number of features is increased as more samples become available. The method of (Raghavan, Madani, and Jones 2006) directly incorporates human feedback in the feature selection procedure through feature weighting, while (Rashidi and Cook 2011) introduce a method that reduces the effort needed for labeling by selecting, at each iteration, all samples matching a rule.

Copyright © 2014, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Active Informative Projection Recovery

Given a dataset $\{(x_1, y_1), \ldots, (x_n, y_n)\} \in \{\mathcal{X} \times \{0, 1\}\}^n$, a RIPR model contains a set Π of subspaces of $\mathcal{X}$, where each subspace is a group of up to d features. A classifier τ_i from the hypothesis space $\mathcal{T}$ is trained on each projection π_i. The model also contains a selection function g which matches a point $x \in \mathcal{X}$ to the projection/discriminator pair with which this point will be classified. The dimensionality d of the subspaces on which the classifiers in the ensemble are trained depends on the application. Typically $d \leq 3$ because of the requirement that human users must understand the classification process. The notation $\pi(x)$ refers to the projection of the point x onto the subspace π while $\tau(\pi(x))$ represents the predicted label for x. The RIPR model class is:

$$\begin{aligned}\mathcal{M} = \{&\Pi = \{\pi;\ \pi \in \mathbf{\Pi}, |\pi| \leq d\},\\ &\tau = \{\tau;\ \tau_i \in \mathcal{T}, \tau_i : \pi_i(\mathcal{X}) \to \mathcal{Y} \quad \forall i = 1 \ldots |\Pi|\},\\ &g \in \{f : \mathcal{X} \to \{1 \ldots |\Pi|\}\}\ \}\ . \end{aligned} \tag{1}$$

Active learning iteratively selects samples for labeling until the model meets some accuracy criteria. Assume that, at iteration k, the samples X_ℓ^k are labeled as Y_ℓ^k and the samples X_u^k are available for labeling. Also let the RIPR model built so far be M^k, with its components Π^k, τ^k and g^k. The problem of selecting samples for informative projection recovery is reduced to finding a scoring function $s : \mathcal{M} \times \mathcal{X} \to \mathbb{R}$, used to select the next sample to be labeled:

$$x^{k+1} = \arg\min_{x \in X_u^k} s(M^k, x)$$

The expected error of a model $M^k = \{\Pi^k, \tau^k, g^k\}$ is

$$Err(M^k) = \mathbb{E}_{x \in \mathcal{X}}[I(\tau^k_{g^k(x)}(\pi^k_{g^k(x)}(x)) \neq y)]$$

We use the notation M_s^k to refer to a model obtained after k iterations of labeling, using the scoring function s. The

aim is to pick the query samples adequately, such that the training error decreases (or at least not increases) with each iteration: $Err(M_s^{k+1}) \leq Err(M_s^k)$. Given the maximum acceptable error ϵ, and a set $\mathcal{S}$ of scoring functions, selecting the optimal strategy can be expressed as follows:

$$s^* = \arg\min_{s \in \mathcal{S}} \min_k \{k \text{ s.t. } Err(M_s^k) \leq \epsilon\} \quad (2)$$

ActiveRIPR starts by requesting labels of a set of r_0 randomly selected samples, building a RIPR model. The standard selection criteria are adapted to RIPR models. The uncertainty sampling score simply considers the lowest conditional entropy on the projections of the model M_{uncrt}^k:

$$s_{uncrt}(x) = \min_{\pi \in \Pi_{uncrt}^k, \tau \in \tau^k} \hat{h}(\tau(\pi(x))|\pi(x)) \quad (3)$$

where $\hat{h}$ denotes the conditional entropy estimator for a label given a subset of the features and $\hat{y}(x)$ is the prediction:

$$g^k(x) := \arg\min_{(\pi,\tau) \in (\Pi^k, \tau^k)} \hat{h}(\tau(\pi(x))|\pi(x)); \hat{y}(x) \quad := \tau^k_{g^k(x)}(x)$$

Query by committee M_{qbc}^k is simply performed by comparing the labels assigned by each of the classifiers in τ_{qbc}^k:

$$s_{qbc}(x) = \max_{\tau_i, \tau_j \in \tau_{qbc}^k} I(\tau_i(\pi_i(x)) \neq \tau_j(\pi_j(x))) \quad (4)$$

In order to sample according to the information gain criterion, we use the notation $\hat{H}^k_{X_0,Y_0}(X_1)$ to represent the estimated conditional entropy of the samples X_1 given the samples X_0 and their labels Y_0. Let $X^k_{\ell,ig}$ be the labeled samples by and $X^k_{u,ig}$ the unlabeled ones. The InfoGain score is:

$$\forall x \in X^k_{u,ig}, \quad s_{ig}(x) = \hat{H}^k_{X_\ell, Y_\ell}(X^k_{u,ig}) - p_0 \hat{H}^k_{X_\ell \cup \{x\}, Y_\ell \cup \{0\}}(X^k_{u,ig}) - p_1 \hat{H}^k_{X_\ell \cup \{x\}, Y_\ell \cup \{1\}}(X^k_{u,ig})$$

Selecting samples with high uncertainty makes sense when there are aspects of the data not yet learned, however, once the informative projections have been discovered, selecting samples with high uncertainty often leads to the selection of purely noisy samples. In this case, selecting the data for which the classification is the most confident improves the model. The score for this query selection criteria is simply the opposite of the uncertainty sampling score:

$$s_{mc}(x) = 1 - \min_{\pi \in \Pi^k_{mc}, \tau \in \tau^k_{mc}} \hat{h}(\tau(\pi(x))|\pi(x)).$$

Experimental Results

The set of synthetic data used in our experiments has 10 features and contains $q = 3$ batches of data points, each made classifiable with high accuracy using one of the available 2-dimensional subspaces (x^1_k, x^2_k) with $k \in \{1 \dots q\}$. The data in batch k also have the property that $x^1_k > t_k$, where t_k is a constant. We also add points that cannot be classified using any low-dimensional models as their labels are assigned randomly. The curves in Figure 1 are averaged over twenty executions of the algorithm. The InfoGain score consistently outperforms the alternatives, followed by the MC score.

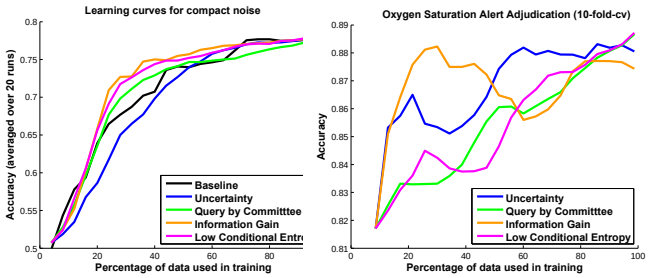

Figure 1: ActiveRIPR on artificial and real data.

We have also used ActiveRIPR to filter out alerts coming from a cardio-respiratory monitoring system, which collects multiple vital signs indicative of the health status of patients. The alerts, each associated to a vital sign, are issued whenever some form of instability requires attention, however, a substantial fraction of these are triggered by malfunctions or inaccuracies of the sensing equipment (artifacts). We extracted a total of 50 features, 800 labeled samples and roughly 8000 unlabeled ones. We use ActiveRIPR to classify oxygen saturation alerts, treating the existing labeled data as the pool of samples available for active learning. We performed 10-fold cross validation, training the ActiveRIPR model on 90% of the samples and using the remainder to calculate the learning curve shown in Figure 1 (right). InfoGain once again outperforms the rest, with accuracy of 0.88 achievable by labeling less than 25% of the total samples. The accuracy decreases because the useful samples have been expended and the samples are not, in fact, iid.

We also used ActiveRIPR with InfoGain sampling to classify blood pressure alerts. This time, we used other classifiers with uncertainty sampling. The table presents the mean leave-one-out accuracy of after 20, 50 and 75 labels. ActiveRIPR performs well, despite using only 6 features.

Samples	K-nn	K-SVM	RF	ActiveRIPR
20	0.61	0.64	**0.65**	**0.65**
50	0.58	0.66	**0.71**	0.70
75	0.6	0.63	0.71	**0.75**

Table 1: Active learning for blood pressure alerts

To summarize, we introduced ActiveRIPR, a method for active learning with feature selection, which improves the learning rates of clinical alert classification while maintaining compact, user-friendly models.

References

Bilgic, M. 2012. Combining active learning and dynamic dimensionality reduction. In *SDM*, 696–707.

Fiterau, M., and Dubrawski, A. 2012. Projection retrieval for classification. In *Advances in Neural Information Processing Systems 25 (NIPS)*, 3032–3040.

Raghavan, H.; Madani, O.; and Jones, R. 2006. Active learning with feedback on features and instances. *Journal of Machine Learning Research* 7:1655–1686.

Rashidi, P., and Cook, D. J. 2011. Ask me better questions: active learning queries based on rule induction. In Apt, C.; Ghosh, J.; and Smyth, P., eds., *KDD*, 904–912. ACM.