Conformal Prediction Based Active Learning by Linear Regression Optimization

Sergio Matiz Romero and Kenneth Barner
Electrical and Computer Engineering University of Delaware, Newark Delaware

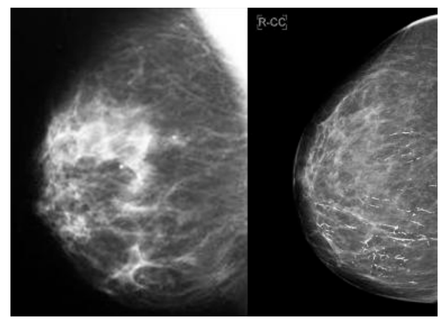
Introduction

- Conformal prediction (CP) uses the degree of strangeness (nonconformity) of new data examples to determine the confidence values of new predictions.
- Conformal predictors can be implemented in conjunction with any traditional pattern classification algorithm, yielding a set of predicted class labels with guaranteed error rate

Motivation

CP can be used in a variety of applications, including:

- Robust face recognition
- Breast cancer diagnosis
- Active learning



Background: Conformal prediction

Let $\{z_1, \dots, z_n\}$ be a bag of elements $z_i = (\mathbf{x}_i, h_i)$. Where $\mathbf{x}_i$ represents a feature vector and h_i its class label.

- **Nonconformity measure:** Function that produces a *nonconformity score* α, which measures how much an instance $\mathbf{x}_{n+j}$ conforms to a particular class h.

- **Example**: Neural networks / SVMs

$$\alpha_{n+j}^{(\mathcal{H}_q)} = -o_{n+j}^{(q)} + \max_{i=1,\dots,M, i\neq q} o_{n+j}^{(i)},$$

- where $o_{n+j}^{(i)}$ is the i-th output of the neural network due to the input $\mathbf{x}_i$

- **p-values:** Represent the probability that $\mathbf{x}_{n+j}$ belongs to a particular class q

$$p_{\alpha_{n+j}}^{(\mathcal{H}_q)} = \frac{count\{i: \alpha_i > \alpha_{n+j}^{(\mathcal{H}_q)}\}}{n+1}, \quad \hat{h}_{n+j} = \operatorname{argmax}_q \, p_{\alpha_{n+j}}^{(\mathcal{H}_q)}$$

The ***confidence*** of $\mathbf{x}_{n+j}$ is defined as

$$c(\mathbf{x}_{n+j}) = p_{n+j}^{(1)} - p_{n+j}^{(2)},$$

where $p_{n+j}^{(1)}$ and $p_{n+j}^{(2)}$ are are the largest and second largest p-values for instance $\mathbf{x}_{n+j}$, respectively

Validity property: $\boldsymbol{\Psi}_{n+j}^{\epsilon} = \{i: p_{\alpha_{n+j}}^{(\mathcal{H}_i)} > \epsilon\}$ contains the correct label for $\mathbf{x}_{n+j}$ with probability $(1-\epsilon)$, where $\epsilon \in [0,1]$ is called significance level.

Background: Active Learning and Query Functions

Set of techniques that automatically select the most relevant/useful data instances to train classifiers

- Labeling instances is an expensive task for large databases (we want to select relevant instances)

Query Functions: Functions used to select examples from the unlabeled pool

- **Uncertainty + Kernel Based Diversity (MCLU-KBD):**

$$\mathbf{x}_t = \operatorname*{argmin}_{\mathbf{x}_i \in T_i / T_d} \left\{ \rho c(\mathbf{x}_i) + (1-\rho) \max_{\mathbf{x}_j \in T_d} \left[\exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma} \right) \right] \right\}$$

Conformal Prediction Based Active Learning by Linear Regression Optimization (CPAL-LR)

Nonconformity measure:

$$\alpha_{n+j}^{(\mathcal{H}_q)} = -\gamma o_{n+j}^{(q)} + (1-\gamma) \max_{i=1,\dots,M, i\neq q} o_{n+j}^{(i)}$$

Proposed Query function: Considers uncertainty, representativeness , and diversity.

$$\hat{\mathbf{r}} = \arg\min_{\mathbf{r}} \|\mathbf{Qr} - \mathbf{u}\|_2^2 + \lambda \|\mathbf{Dr}\|_2^2$$

$$s.t. \quad 0 \leq \mathbf{r} \leq \mathbf{1}$$

u accounts for uncertainty
Q accounts for diversity
r denotes relevance
D accounts for representativeness

$$\mathbf{Qr} = \begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1N} \\ q_{21} & q_{22} & \cdots & q_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ q_{N1} & q_{N2} & \cdots & q_{NN} \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_N \end{bmatrix}$$

$$q_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{\eta} \right)$$

D incorporates representativeness

$$\mathbf{Dr} = \begin{bmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & d_N \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_N \end{bmatrix}$$

d_i is high in sparsely populated regions

$$\hat{d}_i = \sum_{n}^{k} \left\| \mathbf{x}_i - \mathbf{z}_i^{(n)} \right\|_2^2$$

$$d_i = \hat{d}_i / d_{max}$$

Active Learning Algorithm

1. Find classification rule D using the classifier and prop. train. set
2. Compute the nonconformity scores of the unlabeled pool and calibration
3. Compute the p-values

$$p\left(\alpha_{n+j}^{(\mathcal{H}_q)}\right) = \frac{count\left\{i: \alpha_i > \alpha_{n+j}^{(\mathcal{H}_q)}\right\}}{n+1}$$

4. Compute confidence $c(x_{n+j})$
5. Select instances using the proposed query function
6. Return $T_{AL} = T_{prop} \cup T_d$, where T_d contains the N_{AL} instances selected using the proposed query function

Experimental Results

CPAL-LR is evaluated on three databases: The Extended YaleB, AR, and the Caltech101. Example images can be seen in Fig 1.

a) b) c)

Figure 1. Example Images **a)** YaleB **b)** AR **c)** Caltech101

Classification accuracy

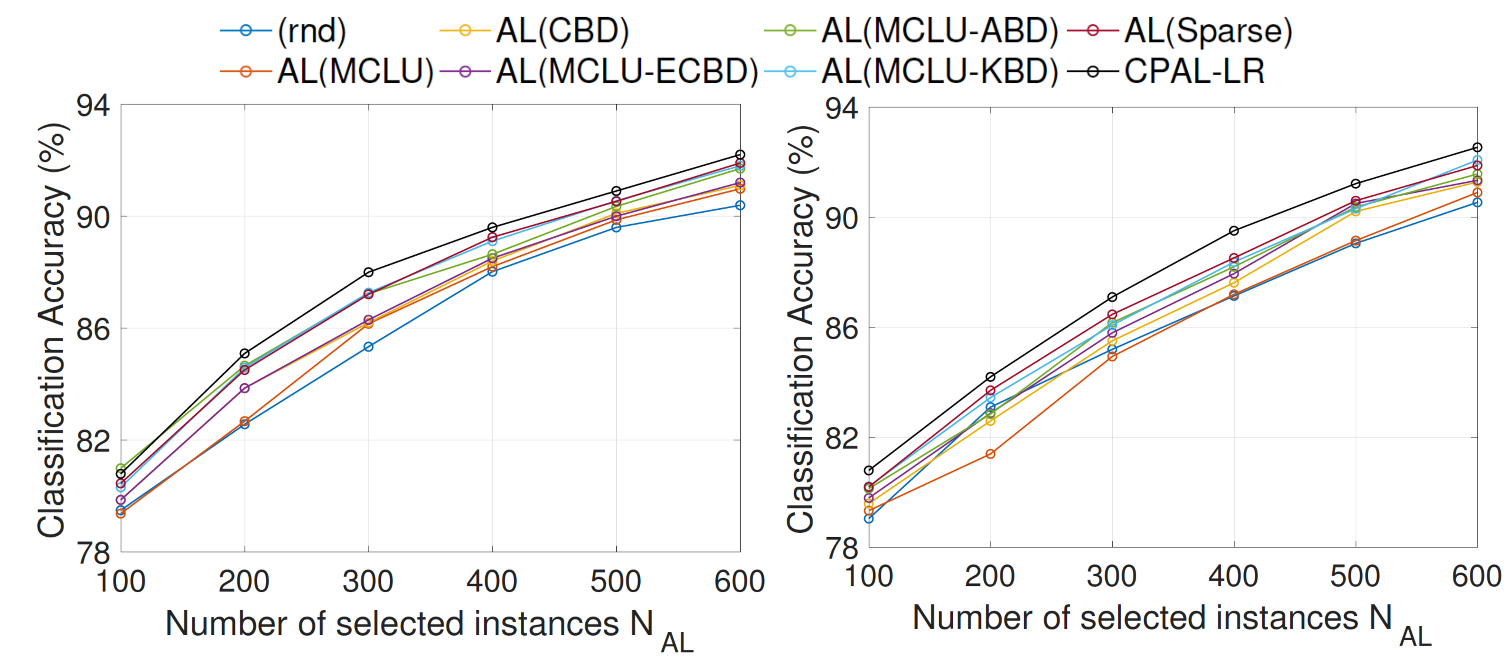


Figure 2. Classification accuracy (%) **a)** YaleB **b)** AR

Algorithm	Query function	YaleB			AR			Caltech101		
		300	400	500	300	400	500	200	300	400
SVM	(rnd)	85.3	88.0	89.6	85.2	87.1	89.0	70.5	71.9	73.2
	AL(MCLU)	86.2	88.2	89.9	84.9	87.2	89.1	71.0	74.2	74.6
	AL(CBD)	86.2	88.4	90.1	85.5	87.6	90.2	71.2	74.2	74.6
	AL(MCLU-ECBD)	86.3	88.5	90.0	85.8	87.9	90.5	71.4	74.5	74.8
	AL(MCLU-ABD)	87.2	88.6	90.4	86.2	88.2	90.4	72.8	74.9	76.6
	AL(MCLU-KBD)	87.3	89.1	90.5	86.1	88.4	90.3	71.9	75.0	76.7
	AL(Sparse)	87.2	89.2	90.5	86.4	88.5	90.6	72.6	74.9	77.0
	CPAL-LR	**88.0**	**89.6**	**90.9**	**87.1**	**89.5**	**91.2**	**73.8**	**75.5**	**77.3**

Quality of the Confidence Values

ValE: The percentage of errors measured as the number of times the correct label for instance $\mathbf{x}_{n+j}$ is not in the set $\boldsymbol{\Psi}_{n+j}^{\epsilon}$

SinP: The percentage of $\boldsymbol{\Psi}_{n+j}^{\epsilon}$ such that $|\boldsymbol{\Psi}_{n+j}^{\epsilon}| = 1$

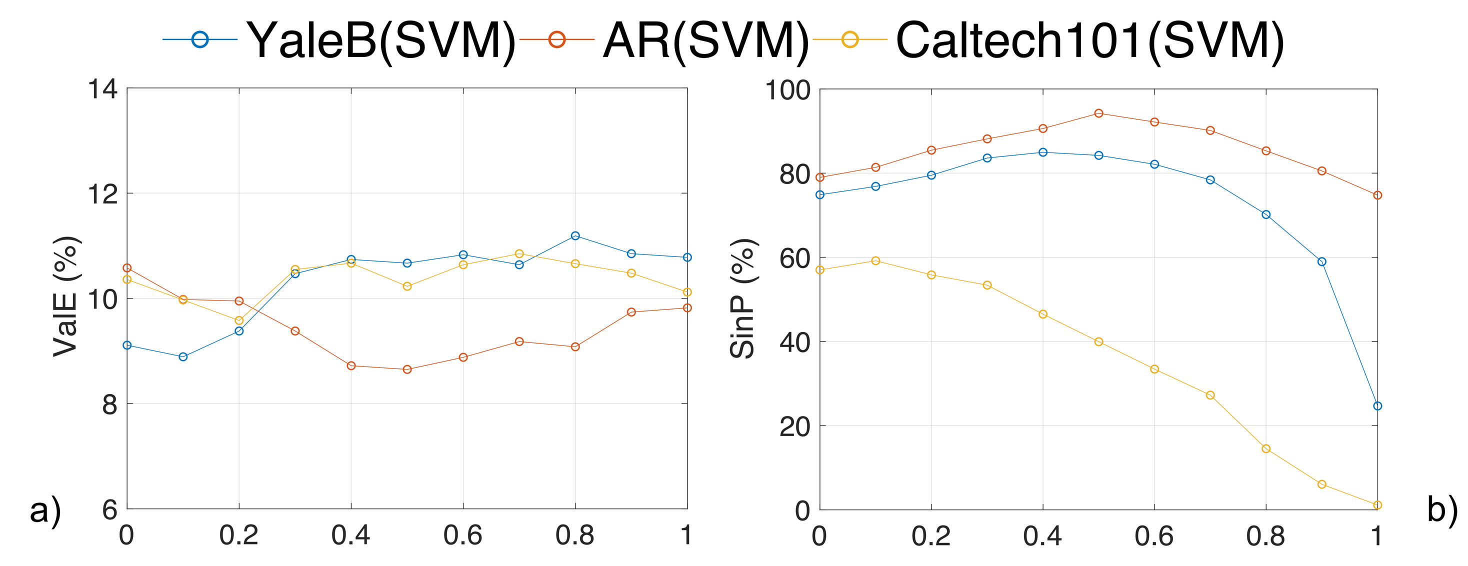


Figure 3. Nonconformity measure performance (%) **a)** ValE **b)** SinP

Conclusion

- CPAL-LR improves the performance of SVMs through active learning, outperforming previously proposed techniques, while producing reliable confidence values

This material is based upon work supported by the National Science Foundation under Grant No. 1319598