Properties of Datasets Predict the Performance of Classifiers

Omid Aghazadeh
http://www.csc.kth.se/~omida
Stefan Carlsson
http://www.csc.kth.se/~stefanc

Computer Vision Group
Computer Vision and Active Perception Laboratory (CVAP)
Royal Institute of Technology (KTH)
Sweden

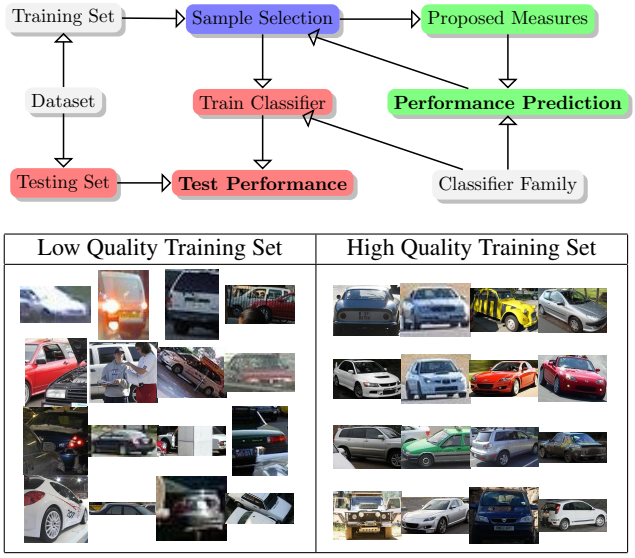

Figure 1: Top: illustration of the proposed procedure. The red boxes comprise the traditional training/testing procedure while the green boxes are proposed in this paper. Bottom: (right) illustration of automatic sample selection (the blue box) using the HOG feature. The low quality set (left) is intentionally generated for comparison. Both set are automatically generated from the "car" class of Pascal VOC 2007, using measures proposed in this paper.

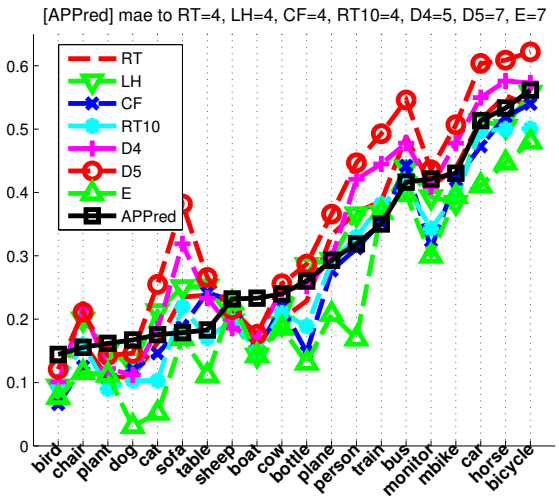

Figure 2: *Test Performance Prediction* of Pascal-VOC 2007 classes (APPred) and the performance of the reference methods. Best viewed electronically and in color.

It has been shown that the performance of classifiers depends not only on the number of training samples, but also on the quality of the training set [5, 6]. The purpose of this paper is to 1) provide quantitative measures that determine the quality of the training set, and 2) provide the relation between the test performance and the proposed measures.

The measures are derived from pairwise affinities between training exemplars of the positive class and they have a generative nature. We use the visual structural similarity measure $K^E_{\text{MMI}}(.,.)$ proposed in [1], which performs feature selection via discriminative reasoning. We aggregate the similarity measures between positive exemplars on local, semi-global and global scales. On each scale we compute the first and second order moments of the quantities in question, and come up with 8 data describing measures.

We show that the performance of the state of the art methods, on the test set, can be reasonably predicted based on the values of the proposed measures on the training set. We assume the training and test sets to be the outcomes of the same underlying distribution and model the test performance as a function of a description of the training set:

$$AP^{(\mathcal{C})}_{\mathcal{M}} = \tilde{f}_{\mathcal{M}}\left(\mu^{(\mathcal{C}_{TR})}\right) + \varepsilon_{\tilde{f}_{\mathcal{M}}} \tag{1}$$

where $AP^{(\mathcal{C})}_{\mathcal{M}}$ is the test performance of the classifier family $\mathcal{M}$ on class $\mathcal{C}$ and $\mu^{(\mathcal{C}_{TR})}$ is a vector describing the training set of the class $\mathcal{C}$. We then assume a sigmoid structure for $\tilde{f}_{\mathcal{M}}$ and model the test performance for a reference set of classifier families $\mathcal{R}$ as

$$\bar{f}_{\mathcal{R}}(\mathbf{w}_{\mathcal{R}};\mathbf{v}) = \left(1+\exp\left\{-\mathbf{w}^T_{\mathcal{R}}\mathbf{v}\right\}\right)^{-1} \tag{2}$$

where $\mathbf{v}$ is a vector describing a training set. Given a data set $\mathcal{D} = \{\mathcal{C}_1,\ldots,\mathcal{C}_D\}$, we solve for $\mathbf{w}^{(\mathcal{C}_{CV})}_{\mathcal{R}} = \arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w},\mathcal{C}_{CV})$ where

$$\mathcal{L}(\mathbf{w},\mathcal{C}_{CV}) = \sum_{\mathcal{M}\in\mathcal{R}} \sum_{\mathcal{C}\in\mathcal{D}\setminus\{\mathcal{C}_{CV}\}} \|AP^{(\mathcal{C})}_{\mathcal{M}} - \bar{f}_{\mathcal{R}}(\mathbf{w};\mathbf{v}^{(\mathcal{C})})\|^2 + \lambda\|\mathbf{w}\|^2 \tag{3}$$

Afterwards, $\mathbf{w}^{(\mathcal{C}_{CV})}_{\mathcal{R}}$ is used to predict the test performance for $\mathcal{C}_{CV}$ and this cross-validating procedure is performed for all $D = 20$ classes of Pascal VOC 2007 [2].

Figure 1 summarizes the proposed procedure. Our approach can be seen as a fundamentally revised version of the earlier approaches that estimate the complexity of the classification problems such as [4]. The results of test performance predictions are shown in figure 2. To summarize the results, we find out that 1) the size of the training set is not a good predictor of the test performance. This essentially means that the 'big data' should meet some quality requirements in order to improve recognition performance. 2) among our measures, the 'connected variation' measure correlates (positively) stronger with the test performances than our measure of 'instra-class variation' (negatively) correlates with them. This suggests that 'big connected data' might rectify the (negative) effects of intra-class variation.

To conclude, this study proposes data-describing measures that link the quality of the training set to the test performance of classifiers. This essentially quantifies the claim on "Unreasonable effectiveness of data" [3] and makes it possible to automatically measure the "cleanness of the data" [6]. This implies that it should be possible to devise rules for the automatic selection of training data that maximize the quality of the training set and consequently increase the test performance.

[1] Omid Aghazadeh, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Mixture component identification and learning for visual recognition. In *European Conference on Computer Vision*, 2012.

[2] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.

[3] Alon Y. Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 2009.

[4] S. Singh. Multiresolution estimates of classification complexity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003.

[5] Antonio Torralba and Alexei A. Efros. Unbiased Look at Dataset Bias. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[6] Xiangxin Zhu, Carl Vondrick, Deva Ramanan, and Charless C. Fowlkes. Do we need more training data or better models for object detection? In *British Machine Vision Conference*, 2012.