Supplementary Materials for "Properties of Datasets Predict the Performance of Classifiers"

Omid Aghazadeh and Stefan Carlsson

Contents

1 Sampling according to Local Connectivity

Figures 1 and 2 depict samples from the training set with different local connectivities (μ_L). This resembles nearest neighbor analysis where nearest neighbors are defined as most similar samples – according to the similarity measure – in contrast to a distance based approach. For each class, the training samples are sorted w.r.t $K_L(p_i) = \max_{p_j \neq p_i} K(p_i, p_j)$ and 16 samples are drawn randomly from the first and the last deciles (10-quantiles) respectively. It can easily be verified that establishing correspondences between poorly connected samples ($\mu_L \approx 0$) is much harder than it is between well connected samples ($\mu_L \approx 1$). This provides qualitative evidence for the importance of connectivity measure in determining the quality of the training set.

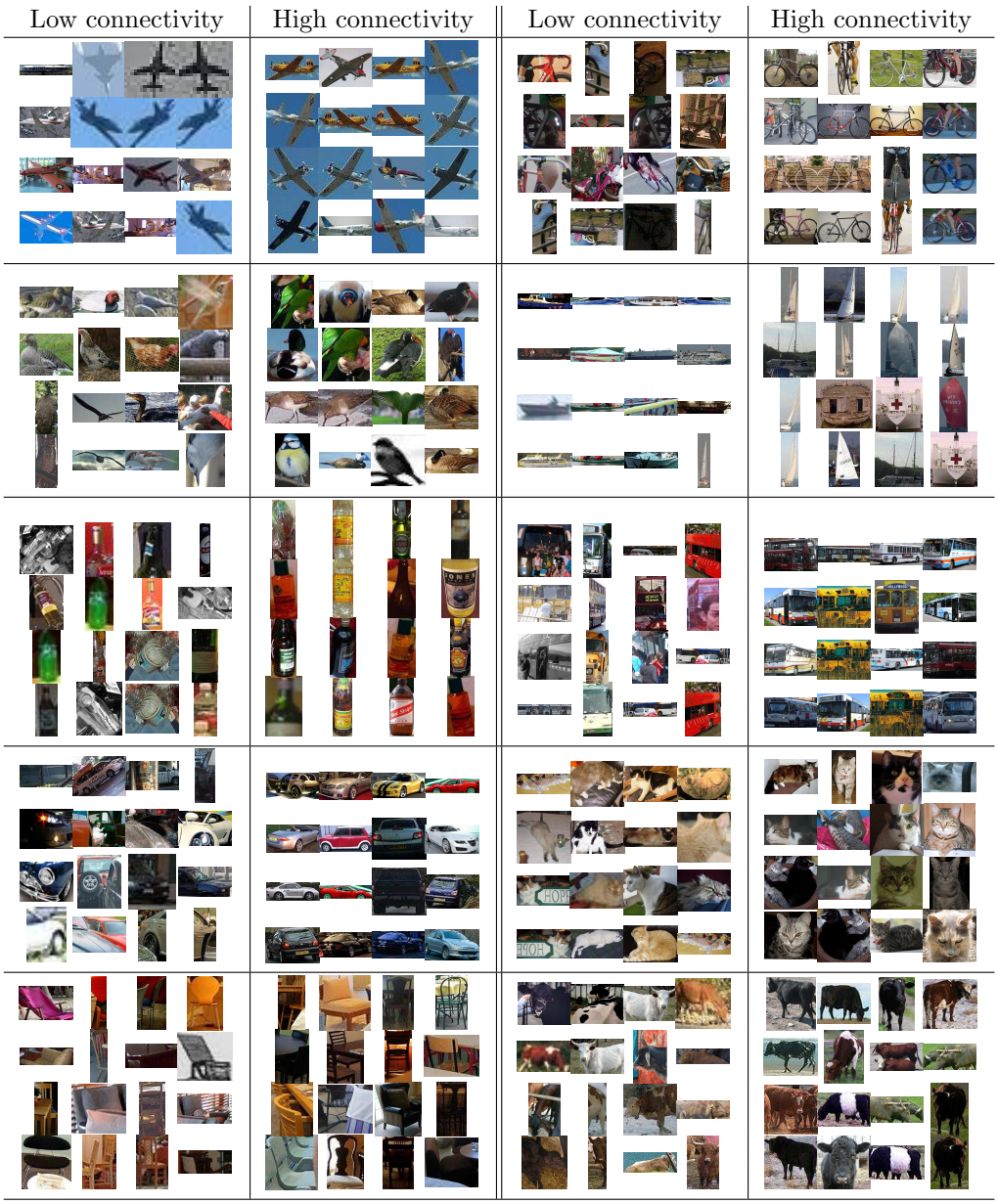

Figure 1: Sampling according to local connectivity (μ_L). For each class, 16 images are randomly sampled from first and last 10-quantiles respectively, according to the local connectivity measure.

Figure 2: Sampling according to local connectivity (μ_L). For each class, 16 images are randomly sampled from first and last 10-quantiles respectively, according to the local connectivity measure.

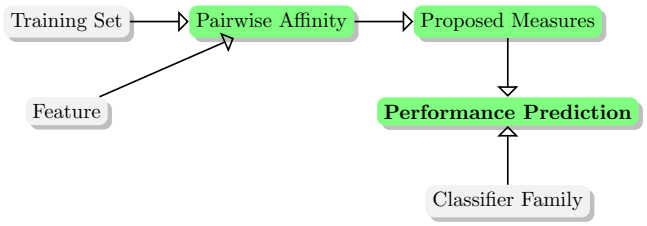

Figure 3: The dependencies between features, classifier families, training data and test performance.

2 Applications / Future Works : Modelling the Interplay Between Features, Classifier Families, Training Data, and Test Performance

In the paper, we mostly focused on analyzing how the test performance varies with properties of the training data while keeping the feature, the similarity measure, the proposed measures and the classifier families fixed. However, the same methodology allows us to model the complete interplay between these factors.

1) Figure 3 demonstrates the dependencies between features, classifier families, training data and test performance. By modelling the entire dependencies at the same time, that is by modelling the predicted performance as a function of all these variables, one could attempt to "optimize" all the variables involved. For example, by varying one or more factors and keeping the rest fixed, one could "optimize" the varying variables (boxes in the figure). Here "optimization" refers to a search process which results in more accurate predictions of test performances. For example, the same proposed procedures can be utilized to select, among a set of **similarity measures**, the one which results in more accurate test performance predictions, while all other factors – the feature, training set, test set, classifier families, and the proposed measures – are kept fixed. As another example, given all the factors but the **feature**, one could select, among a set of possible features, the one which maximizes the predicted test performances, without actually training any classifiers using that feature. Similarly, given a feature, similarity measure,... one should be able to propose the optimal **classifier** [1]. This would be a first systematic approach toward automatic selection of the optimal feature, classifier family, and the **training set**. Hence, this seems the most promising direction to explore further.

2) The assumption that the training set and test set have identical, or at least very similar, distributions is the core assumption of many learning algorithms. It will be interesting to verify to what extent for different classes this core assumption holds by measuring the properties of the test set and comparing it

[1] Automatically proposing the optimal classifier in case of the HOG feature and Pascal VOC 2007 seems not particularly challenging at the moment(see table 1).

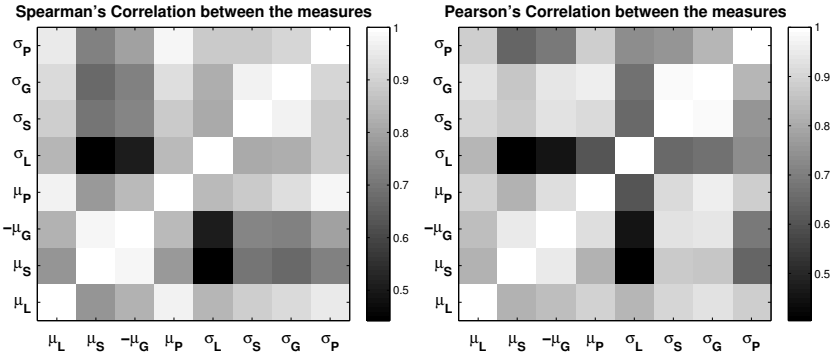

Figure 4: Spearman's Correlation and Pearson's Correlation between the measures. The average Spearman's correlation between the measures is 84.3 while the average Pearson's correlation is 82.9.

to those of the training set. This will automatically determine if a training set is a fair representation of a test set.

Investigating these 2 directions – as discussed in section 4, point 1, in the paper – will make us able to model the part of performance variations that currently our model cannot explain.

3 More experiments and analysis on Pascal-VOC 2007

3.1 Correlation between the moments

Figure 4 shows the Spearman's correlation and Pearson's correlation between the measures. It can be observed that the measures are correlated and that the dependencies are mostly linear. Particularly, the semi-global and global measures seem to be significantly correlated. We provide the following explanation for this observation.

Low global connectivity implies low-length shortest paths, which results in similarity of semi-global and global measures. In the extreme case – where the shortest paths are all of length 1 – global measures and semi-global measures will become the same. This mainly reflects the overall low global connectivity of the Pascal VOC 2007. The strong correlation between local and global connectivity (μ_L and μ_P), (μ_S and μ_G) and (σ_S and σ_G) supports this hypothesis.

3.2 More About Reference Methods

Table 1 shows the test performance of the reference methods and figure 5 shows the correlation between their test performances.

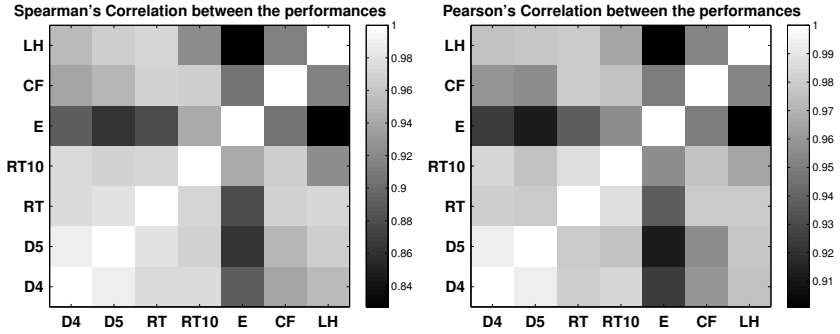

Figure 5: Correlations between test performance of reference methods. The average Spearman's correlation is 0.948 while the average Pearson's correlation is 0.967.

	plane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	monitor	mAP
D4	30	57	10	17	25	48	55	18	22	25	23	11	58	48	42	12	19	32	45	41	32
D5	37	62	12	18	29	55	60	26	21	26	27	15	61	51	45	14	22	38	49	44	**35**
RT	33	54	10	16	23	49	52	16	16	20	24	11	55	44	37	11	23	24	39	41	30
RT10	29	50	10	15	19	41	50	10	16	21	17	10	50	40	33	9	20	22	38	34	27
E	21	48	8	14	13	40	41	5	12	19	11	3	45	39	17	11	23	17	37	30	23
CF	28	54	7	15	15	44	47	15	13	22	24	12	52	42	31	11	23	19	35	31	27
LHSL	29	56	9	14	29	44	51	21	20	19	25	13	50	38	37	15	20	25	37	39	30

Table 1: Test performance of the reference methods. Results are rounded off for better readability.