Feature Selection for Stylometry

Yuxin Fang[1], Yichu He[1] and Yuan Yao[1]

[1] School of Mathematical Sciences, Peking University

Introduction

We authentcate Raphael paintings from forgeries. The data consists of 35 images, of which 13 are works of Raphael, 16 identified as fakes or copies by others, and 6 has their genuineness still under dispution. Previous works proposed to handle it as an outlier detection problem, viewing genuine paintings as samples from the same distributions and forgeries as outliers[**?**]. Feature vectors are used to represent images and variables are selected by forward selection to allow most concentration of genuine works and dispersion of forgeries. While this method achieved high accuracy in leave-one-out validation with van Gogh works, it does not work equally well for the Raphael dataset. In addition, the Raphael dataset contains images whose genuineness is hard to determine. While discarding them is a waste of precious data, it seems inapropriate to put them into either catagory.

In this paper, we address both problems. We mainly focus on variable selection. With features extracted using the same method, we perform stepwise variable selection with an upper bound and a lower bound for variable number. We also proposed a possible way to make use of the disputed data. In addition, we tried using elastic net to avoid problems brought by high correlation and dimension of data. Models are compared for stability of selected variable set, classification accuracy and ability to detect forgeries.

Feature Extraction

Feature extraction is done by using tight frame filtering[**?**].We use a group of 18 3×3 geometric tight frames $\tau_0, \tau_1, ..., \tau_{17}$, consisting of low-pass filter, Sober operators and second-order difference operators in different directions. For P_i, the gray scale intnsity of the i-th image, we convolve it with each filter τ_j, and correspondingly obtain j coefficient matrices of size $m_i \times n_i$:

$$A^{(i,j)} = \{a^{(i,j)}_{m,n}\}_{m_i \times n_i} = P_i * \tau_i, i = 1, 2, ...28; \quad j = 0, 1, ..., 17. \tag{1}$$

Then for each $A^{(i,j)}$, we compute three statistic-s:

- Average value of all entries $\mu_{(i,j)} = \frac{1}{m_i n_i} \sum_{m=1}^{m_i} \sum_{n=1}^{n_i} a^{(i,j)}_{m,n}$
- Standard deviation $\sigma(i,j) = \left(\frac{1}{m_i n_i - 1} \sum_{m=1}^{m_i} \sum_{n=1}^{n_i} (a^{(i,j)}_{m,n} - \mu_{(i,j)})^2 \right)$;
- Percentage of the tail entries $p^{(i,j)} = \frac{\|\hat{A}^{(i,j)}\|_0}{m_i n_i}$, where $\hat{A}^{(i,j)} = \{\hat{a}^{(i,j)}_{m,n}\}_{m_i n_i}$,

$$\hat{a}^{(i,j)}_{m,n} = \left\{ \begin{array}{ll} 1, & \text{if } \ |a^{(i,j)}_{m,n} - \mu^{(i,j)}| > \sigma^{(i,j)}; \\ 0, & \text{otherwise,} \end{array} \right\}$$

and $\|\hat{A}^{(i,j)}\|_0$ denotes of the number of non-zero entries in $\hat{A}^{(i,j)}$.

Thus, for each $i \in \{1, 2, ..., 28\}$, we have a feature vector for the i-th painting:

$$\mathbf{x}_i = \left(\mu^{(i,0)}, ..., \mu^{(i,17)}, \sigma^{(i,0)}, ..., \sigma^{(i,17)}, p^{(i,0)}, ..., p^{(i,17)} \right) \in \mathbb{R}^{54}. \tag{2}$$

Feature Selection

We have obtained a 54 dimensional vector for each image. In this part, we define the outlier detection problem and describe the stage-wise variable selection aiming at maximizing AUC of the ROC curve.

Outlier Detection and Variable Selection Target The forgeries are distinguisheed from genuine paintings by an outlier dection procedure. The assumption is that certain features of Raphael's paintings has simular characteristics that makes them concentrate near a center, while the fogeries are mostly spread far from the center. For the variable selection part, we use only the paintings whose authorship have been determined. That makes a set of 29 paintings.

The goal for the feature selection process is to select a small variable set from the 54 features that best concentrates the genuine paintings and disperse the fakes, and we do this by maximizing the area under the receive operating characteristic (ROC) curve (AUC of ROC). For any feature subset $\mathcal{F} = \{f_1, f_2, ..., f_{|\mathcal{F}|}\}$ with the number of elements $|\mathcal{F}|$, its corresponding feature vector is $\hat{\mathbf{x}}_{i,\mathcal{F}} = (\hat{x}_{i,f_1}, \hat{x}_{i,f_2}, ..., \hat{x}_{i,f_{|\mathcal{F}|}})$, then we use

$$d_i^{\mathcal{F}} = \|\hat{\mathbf{x}}_{i,\mathcal{F}} - \frac{1}{|\mathcal{T}_R|} \sum_{j \in \mathcal{T}_R} \hat{\mathbf{x}}_{j,\mathcal{F}}\|_2, \quad i = 1, 2, ..., 29$$

to obtain the ROC curve w.r.t $\mathcal{F}$ by plotting true positive versus false positive rate for different classifier ρ. Specifically, we sort $\{d_i^{\mathcal{F}}\}_{i=1}^{29}$, and for any number ρ, we can use it as a binary classifier. We want the best feature subset $\mathcal{F}$ to maximize the area under the ROC curve $AUC(\mathcal{F})$.

In computation, we calculate AUC by calculating the p-value in the related Wilcoxon-Mann-Witney test[**?**]:

$$AUC = \frac{\sum_{x_i \in \mathcal{S}_{\text{positive}}} r_i - M(M+1)/2}{MK}$$

, where $\mathcal{S}_{\text{positive}}$ is the set of positive samples, r_i is the rank of d_i and M and K are the number of positive and negative samples respectively.The equivalence of the probability to AUC has been proved in [**?**].

Stage-wise Algorithm We apply greedy stage-wise algorithm to select features.

- *Forward Stage-wise Selection (original method in [**?**])*
  1. Start from $\mathcal{F}_0 = \emptyset$
  2. Choose the next feature: $f_{j+1} = \arg\max_{f \in \mathcal{F}_j^C} AUC(\mathcal{F}_j \cup \{f\})$
  3. Update new subset and return to 2: $\mathcal{F}_{j+1} = \mathcal{F}_j \cup \{f_{j+1}\}$
  4. Terminate when AUC ceases to increase.
- *Stepwise Selection (with both addings and deletions)*
  1. Start from $\mathcal{F}_0 = \emptyset$
  2. Choose a feature from the unchosen set: $f_{j+1} = \arg\max_{f \in \mathcal{F}_j^C} AUC(\mathcal{F}_j \cup \{f\})$
  3. Choose a feature from the chosen set: $f'_{j+1} = \arg\max_{f \in \mathcal{F}_j} AUC(\mathcal{F}_j \backslash \{f\})$
  4. Compare the optimized AUC in 2 and 3 and update the chosen subset with the one with higher AUC; return to 2.
  5. Terminate when AUC ceases to increase.

Make Use of Disputed Data In determining the threshold ρ that separates concentrated data and outliers, we estimate the range by maximizing the accuracy on training set. The accuracy $Acc = (TP + TN)/(TP + TN + FP + FN)$ should be piece-wise constant to ρ, with break points at $\{d_i^{\mathcal{F}}\}_{i=1}^{29}$, thus we obtain a series of intervals $\{I_k = (d_{n_k - 1}, d_{n_k}]\}$. Then we determine the exact value by minimizing the summed distance of disputed images and the boundary:

$$\begin{array}{ll} \min\limits_{\rho} & \sum\limits_{i=1}^{T} (d'_i - \rho)^2 \\ \text{s.t.} & \rho \in I_k \text{ for some } k. \end{array}$$

Other Alternations The greedy stage-wise method is easily disturbed by outliers among the cooncentration set and may turn out to select inappropriate number of features due to some outliers. Therefore, we set the upper and lower limit u and l for number of variables in the stepwise method. Addition to the set is banned when $|\mathcal{F}| \geqslant u$ and deletion of $\mathcal{F}$ can only be done when $|\mathcal{F}| > l$. In addition, the algorithm does not terminate unless $l \leqslant |\mathcal{F}| \leqslant u$.

Experiments

As data is scarce, we did leave-one-out cross-validation on the dataset with certain authorship to evaluate the models.By comparison, we choose $u = 10$ and $l = 4$.

Accuracy

Method	Train Set Accuracy (Mean)	Validation Set Accuracy
Forward	0.6552	0.6207
Stepwise	0.7218	0.6897
Stepwise with Limits	0.7658	0.7586

Stability The size of variable set chosen by the stepwise method seems more stable than that of forward selection, while there does not seems to be much difference between the stability of the variables chosen.

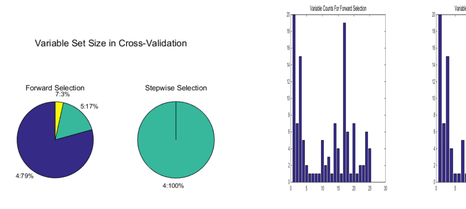

FIGURE 1: Variable set size.

FIGURE 2: Forward selection variable counts.

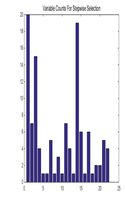

FIGURE 3: Stepwise selection variable counts.

Selected Variables

Frequency	Statistic	Filter	Frequency	Statistic	Filter
20	Mean	τ_0	4	Percentage of Tail	τ_{16}
19	Percentage of Tail	τ_6	3	Variance	τ_1
15	Mean	τ_6	2	Percentage of Tail	τ_{12}
7	Mean	τ_5	2	Percentage of Tail	τ_{13}
7	Percentage of Tail	τ_2	1	Mean	τ_8
6	Percentage of Tail	τ_8	1	Mean	τ_{13}
6	Percentage of Tail	τ_{10}	1	Mean	τ_{15}
5	Mean	τ_{14}	1	Variance	τ_3
5	Percentage of Tail	τ_{14}	1	Percentage of Tail	τ_4
4	Mean	τ_7	1	Percentage of Tail	τ_9
4	Percentage of Tail	τ_3	1	Percentage of Tail	τ_{11}

References

[1] Liu H, Chan R H, Yao Y. Geometric Tight Frame based Stylometry for Art Authentication of van Gogh Paintings[J]. Eprint Arxiv, 2014.

[2] Li Y R, Shen L, Dai D Q, et al. Framelet algorithms for de-blurring images corrupted by impulse plus Gaussian noise.[J]. IEEE Transactions on Image Processing, 2011, 20(7):1822-1837.

[3] Mason S J, Graham N E. Areas beneath the relative operating characteristics (ROC) and levels (ROL) curves: statistical significance and interpretation[J]. Quarterly Journal of the Royal Meteorological Society, 2002, 128(584):2145-2166.